NIST-GCR-91-587

# Expert Systems Approach for Spectra-Structure Correlation for Vapor Phase Infrared Spectra

FILE COPY
DO NOT REMOVE

**Peter R. Griffiths**

University of Idaho
Department of Chemistry
Moscow, ID

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Building Fire Research Laboratory
Gaithersburg, MD 20899

U.S. DEPARTMENT OF COMMERCE
Robert A. Mosbacher, Secretary
NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

NIST

NIST-GCR-91-587

# Expert Systems Approach for Spectra-Structure Correlation for Vapor Phase Infrared Spectra

**Peter R. Griffiths**

University of Idaho
Department of Chemistry
Moscow, ID

Final Progress Report
9/1/88-6/30/89

Issued March 1991

NIST Grant No. 60NANB7D0736

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Building Fire Research Laboratory
Gaithersburg, MD 20899

U.S. DEPARTMENT OF COMMERCE
Robert A. Mosbacher, Secretary
NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

## Notice

This report was prepared for the Building and Fire Research Laboratory of the National Institute of Standards and Technology under Grant Number 60NANB7D0736. The statements and conclusions contained in this report are those of the authors and do not necessarily reflect the views of the National Institute of Standards and Technology or the Building and Fire Research Laboratory.

# FINAL PROGRESS REPORT

Grant No.: 60NANB7D0736

Expert System Approach For Spectra-Structure Correlation

For Vapor Phase Infrared Spectra

Period: 9/1/88-6/30/89

Principal Investigator: Peter R. Griffiths

## Introduction

An expert system can be described as a set of rules that objectively determine the answer to a question given a certain amount of information. Expert systems can be quite simplistic (to the point of triviality such as one parameter and one answer) but usually the name is reserved for complicated systems (hundreds of rules) where it is difficult for a single person to comprehend all the rules at once. As the name implies, an expert system attempts to duplicate the ability of an "expert" to intuitively arrive at an answer from a large amount of data. It may be, as is the case for some skilled medical diagnosticians, that the expert cannot even describe the full set of rules used to arrive at the answer. Thus, most automated expert systems include an algorithm for generating its own rules. That is, the expert system can learn.

One field that is suited to the application of expert systems is the interpretation of chemical data for qualitative analysis. For example, it is possible for a skilled spectroscopist to examine a mid-infrared spectrum of a sample and derive a very good estimate of the molecular structure. This work has focused on

the derivation of expert system for automating this process, or more specifically deriving the algorithm for building the rules for the expert system.

The expert system can be examined in two regimes: the structure or arrangement of the questions(or rules) and the actual component questions. The arrangement of the expert system in this work is simply a list of questions about the structure of the molecule in question. For example:

Does it have an alcohol functionality?

Is it aromatic or aliphatic?

How many carbons does it have?

and so on

Some of the questions are simply answered yes or no, whereas others have some quantitative aspect to them. Many expert systems are arranged in a tree structure where the answer to an initial question will direct which questions are asked next. For the most part such a structure is not efficient for the determination of a molecular structure. The questions are not mutually exclusive (for example, a molecule may contain an alcohol functionality and a carbonyl). Obviously some tree structure can be incorporated into the list (for example, there is no need to ask about the aromatic substitution pattern until the unknown's aromaticity is established). This report is mostly concerned with the smaller scale aspects of the expert system, i.e. the actual algorithm for answering the questions. The expert system approach actually simplifies the process of determining the molecular structure from the infrared spectrum. Instead of having to interpret the entire structure at once, one can ask (and answer) simpler questions, such as: Does the molecule have an alcohol functionality? The next question in developing the expert system is deciding how to extract this information from the spectrum. The approach used in this work is a combination of principal component analysis and classification techniques. The principal component analysis is used to reduce the dimensionality of the spectrum

down to 1 or 2. That is, it provides a window for examining the spectra of a number of compounds such that their spatial relationships can be examined. The classification analysis is then used to classify the different types of molecules (e.g. alcohols and non-alcohols). Unknown sample spectra can then be projected into this space and assigned to one of the groups (i.e. alcohol or non-alcohol). Thus, there is some linear transform of the spectrum that can quantitatively answer a molecular structure question. Each question in the list has its own transform, optimized to provide the best separation between the two classifications (presence or absence of the functional group).

This report is concerned with the application of principal component analysis and a number of classification techniques to two specific structure questions. Is the unknown an alcohol or a non-alcohol? Is the unknown aromatic or non-aromatic? Both of these questions should be relatively easy to answer because both of these functionalities have distinctive absorption bands (the alcohols more so than the aromatics).

An expert system that can objectively determine a molecular structure from the infrared spectrum would be of great utility. It would be a useful adjunct to a hyphenated system, such as GC/FT-IR, where it could identify each component separated by the chromatography. This would greatly reduce the workload associated with the identification of the possibly hundreds of components. It would also be helpful in determining if a chromatographic peak were truly a single component or two unresolved peaks. The assigned structure of a mixture "compound" might be non-sensical or the expert system might be able to respond that the "compound" does not fit into the rules that it has learned (e.g. the expert system responds that the unknown does not fall into either the alcohol or non-alcohol class).

It is not likely that the expert system could derive an exact representation of the structure (some questions are not amenable to infrared spectral determination). It is more likely that the expert system can identify broad classes that the unknown molecules belong to (e.g. aromatic alcohols). Such an "incomplete" expert system

would still be quite useful, particularly in conjunction with a library searching technique. The library can be sorted into classes of molecules and the expert system can tell which lists should be searched for a match. This would greatly reduce search times and increase the efficiency of using extremely large libraries (perhaps to the point of using them in real time with a chromatographic technique)

Theory

If one considers a set of infrared spectra of 1000 data points each, one can think of the spectra as 1000-dimensional vectors. That is, the spectra can be represented as points in a 1000-dimensional space (spectral space). Each axis of the space corresponds to a wavenumber. The coordinate of a spectral point along one axis is the absorbance at the corresponding wavenumber. Each spectrum maps to one and only one point and each point maps to one and only one spectrum. If the set of spectra are associated with compounds of two different classes (say alcohols and non-alcohols), the points in the spectral space can be labeled as such. One can infer that because a trained spectroscopist can determine the presence of an alcohol from the spectrum that the same determination can be accomplished by an examination of the spectral space. No information is lost in the mapping to the space. It is simply a different representation of the same information. In fact, if it were humanly possible to examine the 1000-dimensional space, the distinguishing characteristics of the alcohol class might be readily apparent even to the untrained eye. For example, there may be some plane (or 999-dimensional hyper-plane in this case) that separates the alcohols from the non-alcohols. There may be some more complicated pattern to the difference, but the point is that there is some pattern. If that pattern can be identified and quantitated then an objective algorithm can be devised that determines the class of an unknown molecular structure from the position of the spectral point in the spectral space.

The major difficulty in the procedure described above is the high dimensionality of the spectral space. It is difficult or impossible for a human to

envision a high dimension space, and that makes it difficult to devise a set of pattern recognition rules. Furthermore, if the patterns are simple (e.g. classes separated by simple hyper-planes) then the high-dimensionality is redundant and simply an impediment to the solution. For example, for the identifications of alcohols, the absorption band at 3600 $cm^{-1}$ (the OH stretch) is most often used. Thus, a pattern recognition technique based solely on this region can be used (throwing away 99% of the spectrum and reducing the dimensionality of the spectral space down to around 10). In this study, principal component analysis was used to objectively determine the dimensions of major importance. The data can be pretreated so as force the principal component analysis to represent the spectra in 1 or 2 dimensions such that the separation of the two classes is optimized (vide infra).

Principal component analysis (PCA) has been thoroughly discussed elsewhere and will be given only a cursory treatment in this report. The PCA done in this work was accomplished by singular value decomposition (SVD) which was executed by the non-linear iterative partial least squares (NIPLS) algorithm. In order to perform the PCA and determine the window for separating two classes of compounds, one needs to collect a training set of n spectra with representatives from each class. Each spectrum has m measurements (which must correspond to the same wavenumbers for each sample). The training set of spectra are placed in a n by m data matrix, X, where each row is a sample spectrum and each column corresponds to a wavenumber. SVD decomposes X to the product of three matrices

$$X = USV^T + E \qquad (1)$$

where U (n by p) is the matrix of left singular vectors arranged as columns, $V^T$ (p by m) is the matrix of right singular vectors arranged as rows, S (p by p) is a diagonal matrix of singular values and p is the number of component X was decomposed into and E (n by m) is the matrix of residuals. If p is equal to the rank of X then E is null. The set of column vectors in U is orthonormal and the same is true of the row vectors

in $V^T$. The values in S refer to the relative importance (or scale) of the vectors in determining X. The order of the vectors in U and V and the values in S is set up such that left, upper values in S are greater in magnitude than those to the right and below. Thus, the first vector in U and in V describe the most of X (which means that most of the variance is modelled by the first vectors).

The values in U are referred to as scores and the values in V are loadings. There is a row in U for each sample. In fact the rows in U can be thought of as reduced dimensionality spectra (dimension p instead of m). Let us consider the case where p equals two. Each sample will have two scores [s1,s2]. The training set spectra can be represented in a two dimensional scatter plot. Each sample will have two coordinates [s1,s2] which defines its position in the plot. This score plot can be thought of as a two dimensional window (or cross-section) of the spectral space.

Of course p can be greater than 2 and the plot can done with any pair of components. If the first two components are used then the plot represents the cross-section with the greatest width (variance), because the first components model the greatest variance in the matrix. One would hope that the two classes of samples are well separated in one of these scores plots. The two classes might be separable by merely drawing a line between the two groups or by using a Mahalanobis distance or a nearest neighbor approach.

One of the most useful aspects of PCA is that a new, unknown spectrum can easily be projected onto one of these scores plots. For a given component i one merely takes the dot product of the unknown's spectrum and row i from $V^T$ and then divides by element $s_{ii}$ for S. This provides the score for this unknown and this component. By placing the unknown on the plot one can determine the unknown's class by one of several different classification techniques.

As described above, there is one difficulty in looking at the scores plots. One is not guaranteed that the class separation is optimized in the first two component scores, or for that matter in any pair of components. There are a couple of data pretreatments that can be applied to rotate the scores plots towards separating the

two classes.

The first step in the data pretreatment is autoscaling. Autoscaling consists of two steps: mean centering and variance scaling. Mean centering consists merely of determining the mean spectrum of the training set and subtracting that mean spectrum from each member of the training set. If one considers the data set as a collection of points in the spectral space, mean centering moves the collection of points such that the mean is at the origin. Variance scaling consists of determining the variance of each measurement (finding the variance spectrum) across the training set and dividing each measurement in each sample by the square root of the appropriate variance. In this work, the square root of the sum of squares (SQSS) was used instead of the square root of the variance, but the two cases are equivalent, differing only by a constant, (1/SQRT(n-1)), across the spectrum. Graphically, variance scaling is more difficult to interpret (at least in the spectral space). If one considers the sample space, where each axis corresponds to a sample and each measurement corresponds to a vector, then dividing the measurements by the SQSS sets the vectors to unit length.

As far as the PCA is concerned, autoscaling the data has the effect of giving all the measurements equal weight in the determination of the principal components. With the original data, a wavenumber where there is an absorbance maximum has a greater effect on the "shape" of the collection of sample points in the spectral space than a wavenumber in a baseline region. After autoscaling, the magnitudes of the variance for each measurement (size of the collection of sample points in that direction) are all equal. Autoscaling is normally used in cases where the measurements have different units (for example, samples of sea water could be characterized by collection depth and magnesium concentration). To compare distance and concentration fairly in a multivariate technique such as PCA, one autoscales the data which makes them unitless. Measurements are expressed as standard deviations away from the mean. One can make the arguement that autoscaling is inappropriate for spectroscopic data because the measurements that are weighted against are those where

there is a high absorbance which corresponds to a high signal. That is, one is amplifying the baseline at the expense of the signal. While that is true, in this analysis it is not known a priori what the spectral distinction between two classes will be. The difference between two classes may be subtle and that subtle change may be swamped by the apparently random variations in the large peaks in the spectra. Thus after autoscaling all the measurements have an equal chance at showing their ability to distinguish the two classes.

The next step in the pretreatment is feature weighting. Feature weighting scales the equally weighted measurements such that they are no longer equally weighted. The measurements are now weighted not by their original magnitudes but by their ability to separate the two classes. A feature weight $w_k$ is derived for each measurement k. Each measurement is then multiplied by the feature weight. The feature weight is calculated as follows

$$w_k(I,II) = \frac{\Sigma x_I^2/N_I + \Sigma x_{II}^2/N_{II} - 2\Sigma x_I \Sigma x_{II}/N_I N_{II}}{\Sigma(x_I - x_I)^2/N_I + \Sigma(x_{II} - x_{II})^2/N_{II}} \qquad (2)$$

The subscripts I and II refer to the two classes of samples. The feature weight is the ratio of the intercategory variances to the sum of the intracategory variances. The greater the discriminating ability of a particular measurment (wavenumber) the greater the feature weight. If a measurement has no discriminating power then the feature weight is one. The feature weight is analogous to the resolution parameter of chromatography. If one considers a histogram of the values of a measurement for two classes of samples, a discriminatory measurement will produce a bimodal distribution. The feature weight is the ratio of the variance between the two modes and the variance inside each mode.

One can interpret the effect of the feature weighting graphically. In the spectral space the training set samples define a certain shape. The PCA will define the "major axes" of the shape. The feature weighting stretches the shape in the direction of those axes (measurements) where the classes are well separated. In

stretching the shape, the PCA is forced to rotate the principal components towards these directions and thus the separation in the scores plots are improved. The window is turned towards a more efficient direction for the separation. Prior to the feature weighting, the PCA has no information about the two classes. All samples are equivalent. Feature weighting is the mechanism for including classification information in the PCA.

There are many ways to define a weighting scheme. In this work, feature weights as defined above and squared feature weights (which are merely the square of the value described above) were used. Squared feature weights are similar to the original feature weights (obviously) but emphasize the large weights relative to the small weights.

Experimental

For this report two classifications were attempted: alcohols and non-alcohols and aromatics and non-aromatics. The data used were extracted from a library of vapor-phase FTIR spectra supplied by Sadtler Research Laboratories (Division of Bio-Rad Laboratories, Inc. Philadelphia, Penn.). For the alcohol experiment 52 spectra were extracted (26 alcohols and 26 non-alcohols). For the aromatic experiment 50 spectra were extracted (25 aromatics and 25 non-aromatics). The samples were chosen by hand with the intent to provide a wide range of molecular types within each classification. The spectra were converted to ASCII code using Spectra-Calc (Galactic Industries Corp, Salem NH.). Because the PCA program used cannot handle very large amounts of data, the spectra were deresolved by keeping every fourth data point resulting in 460 data points between 4000 and 470 $cm^{-1}$. Five runs of the PCA program were done: (1) untreated alcohol data, (2) autoscaled and feature weighted alcohol data, (3) autoscaled and squared feature weighted alcohol data, (4) autoscaled and feature weighted aromatic data, and (5) autoscaled and squared feature weighted aromatic data. The spectra were deresolved and treated by software written in Turbo Pascal (Borland International, Scotts Valley, Ca.). The PCA was also done

by a program written in Turbo Pascal. The scores plots and other graphs were produced using Lotus 123 (Lotus Development Corp.) and Windows Draw (Microsoft, Bellevue Wa.).

## Results

*Trial 1*

In this trial the PCA was run on the raw spectral data for the alcohols and non-alcohols. None of the scores plots showed a good separation between the two classes. The best separation (which can only be characterized as fair at best) was achieved on the plot of component 3 versus 5 (fig 1) where the o's and x's designate the alcohols and non-alcohols respectively. Figure 2 shows the same scores plot with ellipses drawn around the two classes. The inner ellipses are set so that the radius parallel to an axis is the standard deviation of class scores of that component. The outer ellipses are defined the same way with a student's t value of 2.06 to define the 95% confidence limits. That is, fitting a Gaussian distribution to the scores will put 95% of the distribution area within the outer ellipse. One can see that the inner ellipses overlap each other substantially and within the non-alcohol inner ellipse there are 7 alcohols. The separation shown on this plot is not very good. To be fair, there may be a better separation if one considers more than 2 components at a time. The overlap may be resolved in a third or fourth dimension.

Figures 3 and 4 show the loading vectors plotted in a spectral format for the third and fifth components. Though it must always be remembered that the loadings are only mathematical components of the data, their characteristics can be interpreted in terms of spectra. If a component was successful in discriminating a class of compounds then the corresponding loading must reflect the spectral information inherent in that discrimination (unfortunately the loading can easily include extraneous spectral features). Therefore, when discriminating between alcohols and non-alcohols it is not unreasonable to expect to see the generic alcohol spectrum in the loading vectors. For alcohols one would expect to see a band around 3600 $cm^{-1}$ for

the O-H stretch. For loading 3 a small downward-going feature is observed in the correct region and for loading 5 there is a triplet structure in the region. Otherwise, there is no real indication of alcohol spectral character in these loadings and it is not surprising that the separation done by these components is poor.

Furthermore, the amount of the data explained by the third and fifth components in this analysis is very small. The amount of information in the data set can be quantitated by the Frobenius norm (F-norm). The F-norm is the square root of the sum of squares of the elements in X. As principal components are subtracted from the matrix the F-norm drops and each successive component decreases the F-norm by smaller amounts (because they explain less of the total variance). Figure 5 shows a plot of the F-norm as the components are removed from X. Before any components are removed (i.e. the original data) the F-norm is about 23.5 and the first component drops it to about half that. The succeeding components explain relatively little of the total. Components 3 and 5 represent very little of the original data even though they show the best class separation.

*Trial 2*

To improve the separation autoscaling and feature weighting were performed on the data for trial 2. Figure 6 shows the mean spectrum that was subtracted from the training set spectra and figure 7 shows the SQSS spectrum that was dividing from the training set spectra. The mean and SQSS spectra show small peaks in the O-H stretching region, but both plots are dominated by the C-H stretching band around 3000 $cm^{-1}$. This is the dominating effect that produced the outcome for trial 1. The feature weights are plotted as a spectrum in figure 8. The largest feature weight is at the 3600$cm^{-1}$ region just as one should expect. The peak at 1000-1100 $cm^{-1}$ can be attributed to the C-O stretch. There is quite a bit of structure distributed throughout the rest of the "spectrum". These peaks should not necessarily be interpreted as alcohol absorption bands. These peaks show up because a finite sample set was used. For example, it is

entirely possible that one of the two subsets in the sample set (alcohols or non-alcohols) has more chlorinated compounds than the other. If this is so, then the C-Cl bands will also discriminate between the two classes and will generate a large feature weight. The mathematics have no way of telling whether a band should or should not be used except for its discrimination ability. The way to offset this effect is to use large data sets where the presence of other types of functionalities is mostly uncorrelated with the presence of the alcohol. This is difficult to accomplish perfectly with only 52 samples.

The scores plot for components 1 and 2 (fig. 9) show a good separation between the two classes. Not only is the separation much improved over trial 1 but the best separating power is shown by the first two components instead of the 3rd and 5th. The major variance of the data set is now the difference between the two classes. This is an effect of the stretching of the collection of samples in the spectral space as described above. The same plot with the ellipses (1 standard deviation and 2.06 standard deviations (95%)) is shown in figure 10. Note that the inner ellipses do not overlap and there is only one alcohol in the non-alcohol inner ellipse. The loading vectors (fig. 11 and 12) for these components are more complicated than those for trial 1 (an effect of the autoscaling) but the maximum values in each is at the 3600 $cm^{-1}$ region.

*Trial 3*

The next trial employed the same data pretreatment as trial 2 except that the feature weights were squared. The squared feature weights are plotted as a spectrum in figure 13. As one would expect, the maximum peak (at 3600 $cm^{-1}$) is even larger than before compared to the rest of the weights. Squaring the weights has the effect of placing more of the emphasis on one or more bands giving the maximum separating power. The data set is stretched even more and in the direction of the axis corresponding to 3600 $cm^{-1}$. Squaring the weights has the additional benefit from a spectroscopic viewpoint of discriminating against those smaller peaks in the feature weight

spectrum that correspond to chance correlations that occur in small data sets. One is more assured that the weighted measurements are truly relevant to the classification required. If squaring the weights is effective then one might wonder if cubing or raising to the 10th power would be even more effective. As the power is increased the maximum feature weight is emphasized until it is effectively the only measurement used and one loses the multivariate advantages of the PCA. Thus there is some optimal power to used with each data set in terms of producing the best separation. In this study this parameter was not optimized. The powers of 1 and 2 were both tested.

Squaring the feature weights effected a remarkable improvement in the separation as shown in the scores plot for components 1 and 2 (fig 14). The two classes are well separated except for one outlier alcohol in the middle of the non-alcohol group. This outlier is o-ethoxyphenol. Examining the spectrum of this compound it was found that the O-H stretching band was shifted to lower wavenumber because of the intra-molecular hydrogen bonding of the alcohol hydrogen to the neighboring ether oxygen. Compared to the squared feature weight spectrum (fig 13) this O-H stretch band was shifted into the low weight flat region just below the maximum. It is not surprising that the scores plot classified this compound as a non-alcohol when the plot is based so heavily on the 3600 $cm^{-1}$ region.

One might also not that the non-alcohols are more tightly grouped together than the alcohols. This may seem counter-intuitive because the allowed structural variation is greater for non-alcohols than it is for the compounds constrained to be alcohols. However, this tightness is explainable, as the separation is based mostly on one measurement. Non-alcohols all respond as "0" when the 3600 $cm^{-1}$ region is measured and alcohols respond with a much greater variance (depending on peak intensity and exact location (i.e. overlap with the loadings)).

The loading vectors for components 1 and 2 are plotted as spectra in figures 15 and 16. Both loading vectors are dominated by the peak at 3600 $cm^{-1}$ region. Additionally one might note that the shape of the first loading vector is nearly

identical to that of the squared feature weights (except of course for the shift in aseline as the minimum weight is 1). This similarity is also apparent between the first loading vector and the weights for trial 2 but it is not as marked as in this case. This similarity is not completely surprising because the weights do stretch the data set shape and as the weights become greater the first principal component is forced more in the direction of the stretch.

*Trial 4*

In this trial a new training set was generated for the examination of aromatics and non-aromatics. One might expect that the separation of aromatics from non-aromatics will be a little more difficult than the alcohol case because the characteristic aromatic bands are overlapped with other bands and not isolated as in the alcohol case. In this trial the data were autoscaled and feature weighted (the raw data case was not attempted). The mean and SQSS spectra are plotted in figures 17 and 18 respectively. The feature weights are plotted as a spectrum in figure 19. The maximum weight is only 2.5 (compared to almost 4.5 for the alcohol case). This low weight magnitude is the effect of the overlapped bands. There is no isolated measurement (single wavelength) that is a consistent indicator of aromaticity. The maximum weight peak occurs just above 3000 $cm^{-1}$ which corresponds to the aromatic C-H stretch.

As in the previous feature weighted cases, the first two principal components to the best job of separating the two classes. The scores plot shows that a fair to good separation is achieved (fig. 20). A line is drawn that separates the two classes as efficiently as possible. The slope of the line is determined by projecting the data points onto the orthogonal line and fitting a Gaussian distribution to each class. The separation of the two distributions is calculated by the difference between the class means divided by the sum of the class standard deviations. The slope of the line is adjusted iteratively to optimize the separation parameter using a simplex optimization. After the slope is determined, the intercept is determined such that

the line intersects the orthogonal line where the sum of the two distributions is at a minimum. In this plot, all of the aromatic samples are above the line and all but four of the non-aromatics are below.

The same score plot (fig 21) with the ellipses (1 standard deviation and 2.06 standard deviations) show that the inner ellipses overlap a slight amount. In comparison, the equivalent plot for the alcohols (fig 10) showed well separated inner ellipses.

The loading vectors for these two components (fig 22 and 23) show no interpretable structure other than the peak at the 3100-3000 $cm^{-1}$ region. Of note, is that the peak dominates the second loading vector but not the first. The scores plot shows that the separation between the classes is mostly along the second component, whereas for the alcohols it was mostly along the first. Furthermore, the structure in the first loading is negative and the aromatic samples tend to have negative scores for the first component. The converse is true for the second component (positive structure and positive scores). Taken together this indicates that the structure is present in aromatic spectra in a positive fashion. That is, the separation is dependent on a spectral attribute that is present in aromatics and not in non-aromatics and not the other way around.

*Trial 5*

Analogous to trial 3, the feature weights for the aromatic case were squared in an attempt to improve the separation. Figure 24 shows the squared feature weights plotted as a spectrum. As expected the region between 3100 and 3000 $cm^{-1}$ dominates the weights. The scores plot for the first two principal components (fig 25) shows a number of effects. First of all, the separation is not drastically improved. While the non-aromatics are more tightly grouped, there are still three non-aromatics that appear in the aromatic region (methyl 3-pyridyl ketone, 2-methyl-2-pentene and 2-amino-3-picoline) and one non-aromatic a great distance from the non-aromatic group (acetonitrile) and now there are two aromatics in the non-aromatic region (1,2,4-

trifluorobenzene and 5-chloro-2-nitrotoluene). The plot with the ellipses (fig 26) shows that the inner ellipses still overlap a small amount. Thus squaring the feature weights was not terribly effective in this case. Secondly the main separation is effected along the first axis and not the second. Both loading vectors (fig 27 and 28) show a large peak at the 3100-3000 $cm^{-1}$ region and both are positive. The first loading and the weights are similar but not to the degree exhibited in trial 3. The squared feature weights did move the scores plot window, but the new window was not substantially better. These observations can be explained in light of the difference between the distinguishing characteristics of the alcohols and aromatics. In the alcohols, there is a single distinctive alcohol measurement (3600 $cm^{-1}$) and emphasizing it while losing the rest of the spectrum resulted in a better separation. In the aromatics, there is no single distinctive measurement. The best region (3100-3000 $cm^{-1}$) is heavily overlapped with the aliphatic C-H stretch. To separate the aromatics from non-aromatics requires the multivariate advantages of the PCA. Emphasizing the best region at the expense of the multivariate advantages of using the rest of the spectrum did not produce a better separation.

*Validation Trial*

In any chemometric technique such as the one described here, it is important to do a validation of the model. In this work the model developed in trial 3 was tested by projecting 10 samples (5 alcohols and 5 non-alcohols) into the scores plots. The spectra were extracted from the library and trimmed to the same measurement set. The mean spectrum from the training set was subtracted from each spectrum in the validation set. The SQSS spectrum was also divided from the validation set and the squared feature weights were multiplied in. The dot products of the validation spectra and the first two loading vectors were determined and scaled by dividing by the corresponding singular value. Figure 29 shows these values plotted onto the original training set score plot (see fig 14). The +'s correspond to the validation alcohols and the *'s correspond to the validation non-alcohols. One can see that the

model effectively classifies the validation samples into the correct groups. There is one alcohol sample that lies quite apart from the rest (3-ethylphenol). This suggests that phenols are different enough from aliphatic alcohols that it is not efficient to group the two classes together into one test. That is there should be one test for aliphatic alcohols and a separate test for phenols. The phenol is dropped from the plot in figure 30. One can see that the five non-alcohols are clearly identified. Of the 4 remaining alcohols, 3 are clearly separated from the non-alcohol class and 1 is questionable but is closer to the correct class than the incorrect class. It appears easier to tell that an unknown is a non-alcohol that it is to tell that it is an alcohol. Since these two classes are exclusive the two questions can be combined to provide a confident response.

## Conclusion

An expert system can use a principal components analysis approach to determine the structure classifications of unknown molecules from their IR spectra. The determination of a rule (i.e. setting up a training set, doing the pretreatment and the PCA) takes about one microcomputer-day. Once the rule is determined, the projection and classification of unknowns is very fast. Most current expert systems work on peak tables. That is, they ask whether there is peak in a given region of a certain intensity and width. This approach requires that the spectrum to be decomposed into a peak table which is a time consuming process and sometimes requires the subjective input of the user. The PCA-based approach obviates the need for this step. This PCA-based expert system is a valuable tool for the spectroscopist and analytical chemist.

## Acknowledgements

The authors would like to thank Erik Hasenoehrl for his assistance in extracting the spectra from the Sadtler library and Richard Jackson for his helpful suggestions concerning the classification techniques.

Bibliography

Sharaf M.A., Illman D.I., Kowalski, B.R., Chemometrics, John Wiley & Sons, New York, 1986

Pierce T.H., Hohne B.A., Artificial Intelligence Applications in Chemistry, American Chemical Society, Washington DC, 1986

Nyquist R.A., The Interpretation of Vapor-Phase Infrared Spectra, Volume 1, Group Frequency Data, Sadtler Research Laboratories, Philadelphia, Pa, 1984

Figure 1 Scores plot for components 3 and 5 for the alcohol/non-alcohol classification with untreated data.

# Untreated Data Set

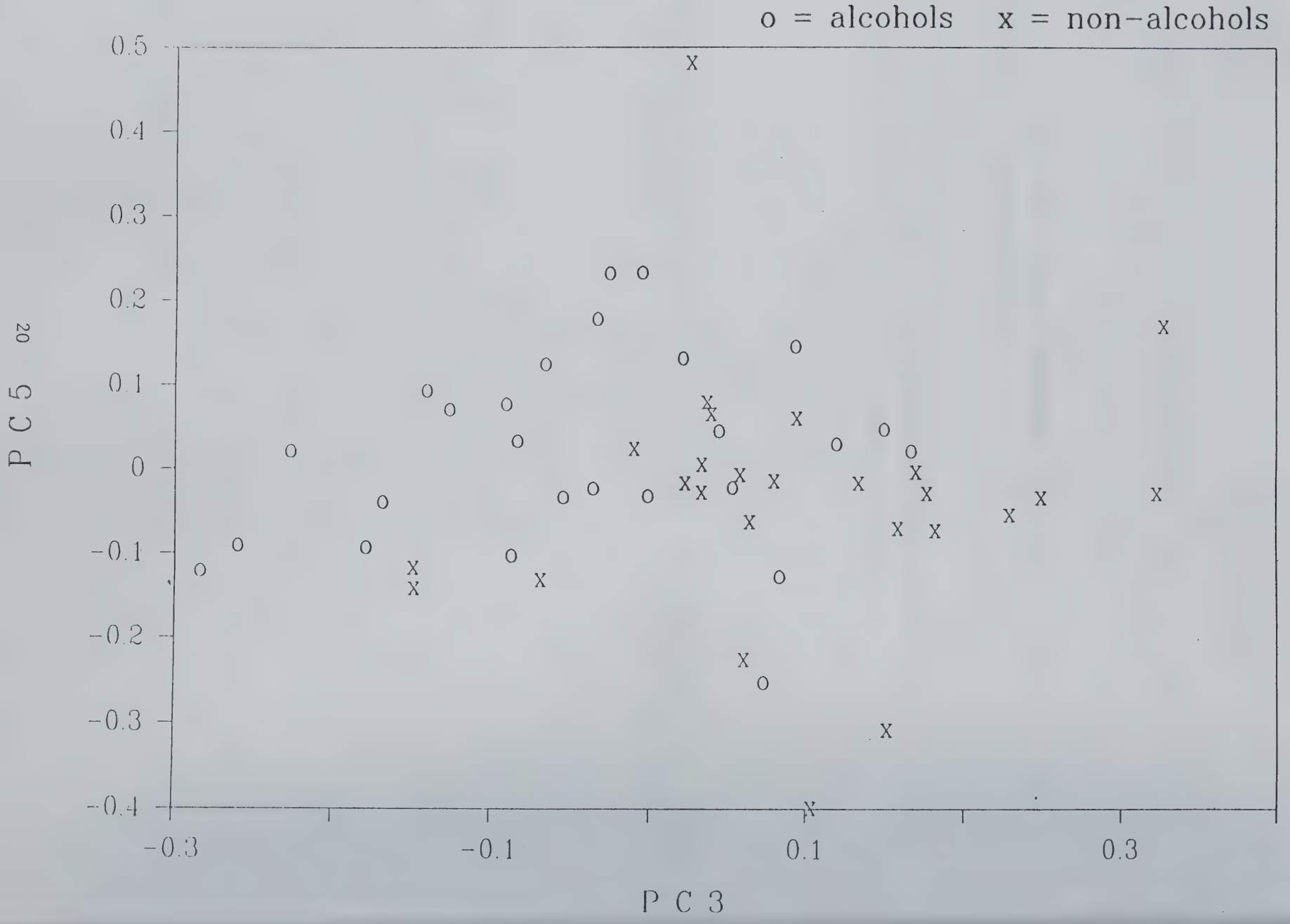


Figure 2 Scores plot for components 3 and 5 for the alcohol/non-alcohol classification with untreated data with ellipses set at 1 standard deviation and 2.06 standard deviations (95% confidence limits).

# Untreated Data Set

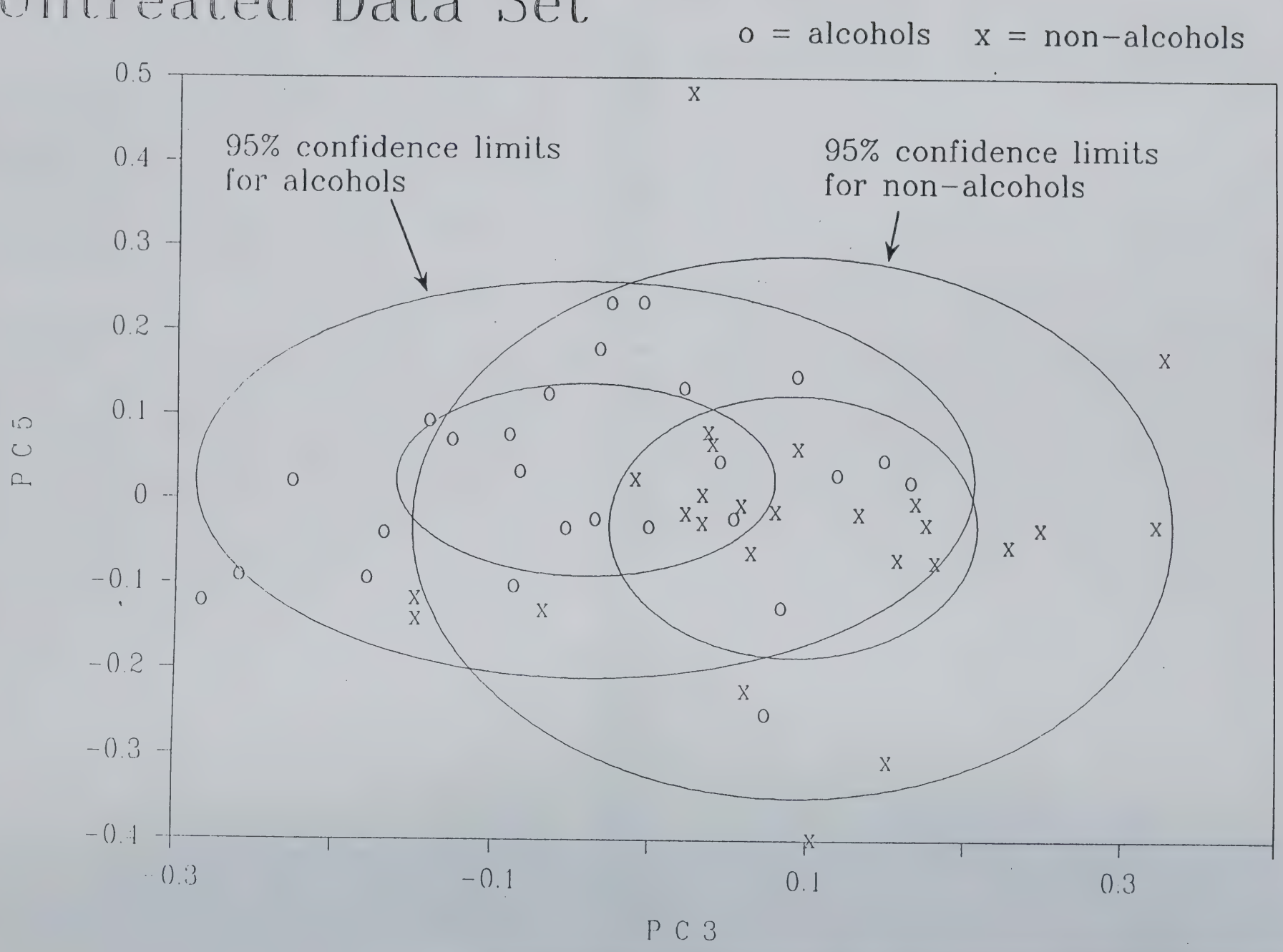


Figure 3 Loading plot for component 3 for the alcohol/non-alcohol classification with untreated data.

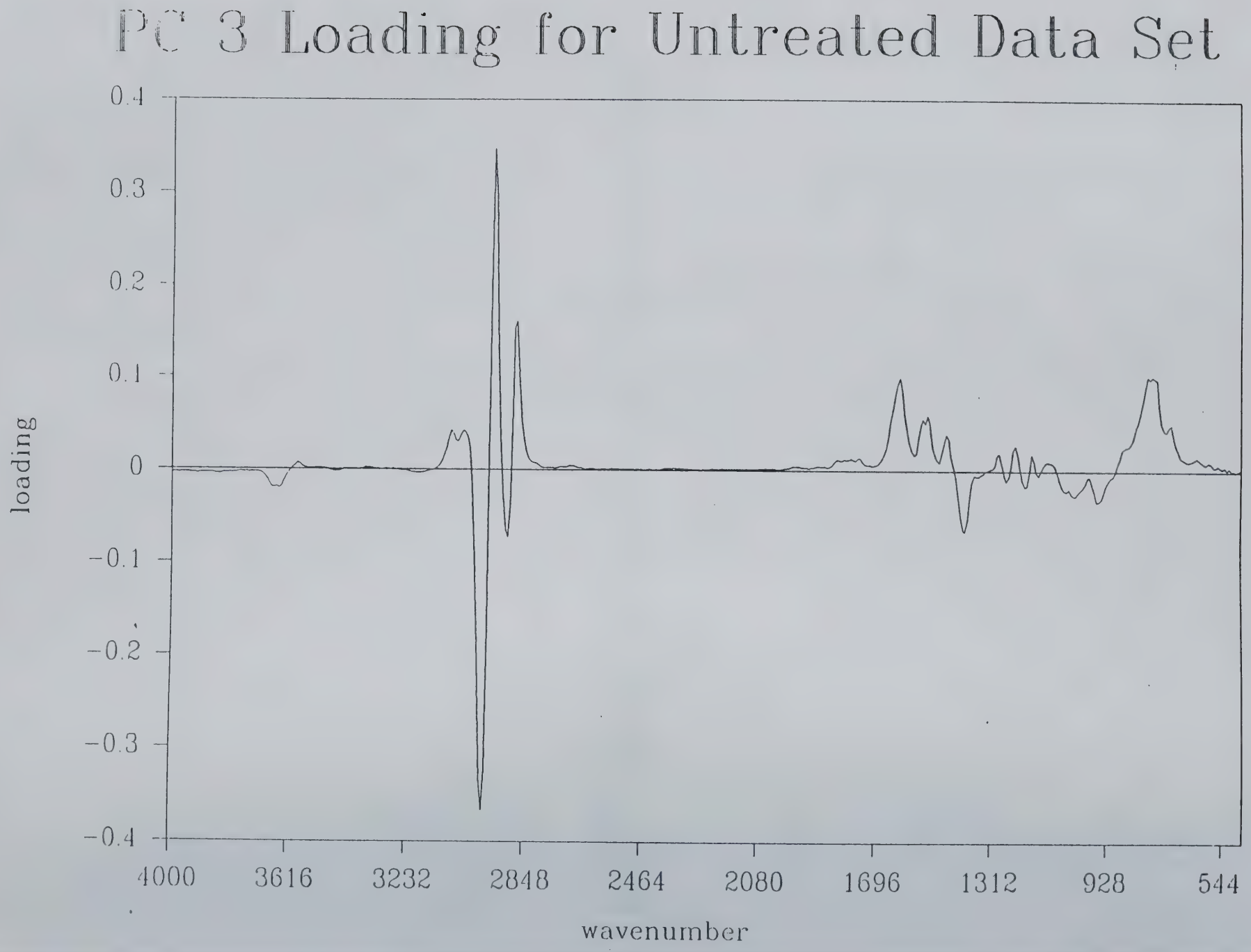

PC 3 Loading for Untreated Data Set
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
loading
4000
3616
3232
2848
2464
2080
1696
1312
928
544
wavenumber

Figure 4 Loading plot for component 5 for the alcohol/non-alcohol classification with untreated data.

# PC 5 Loading for Untreated Data Set

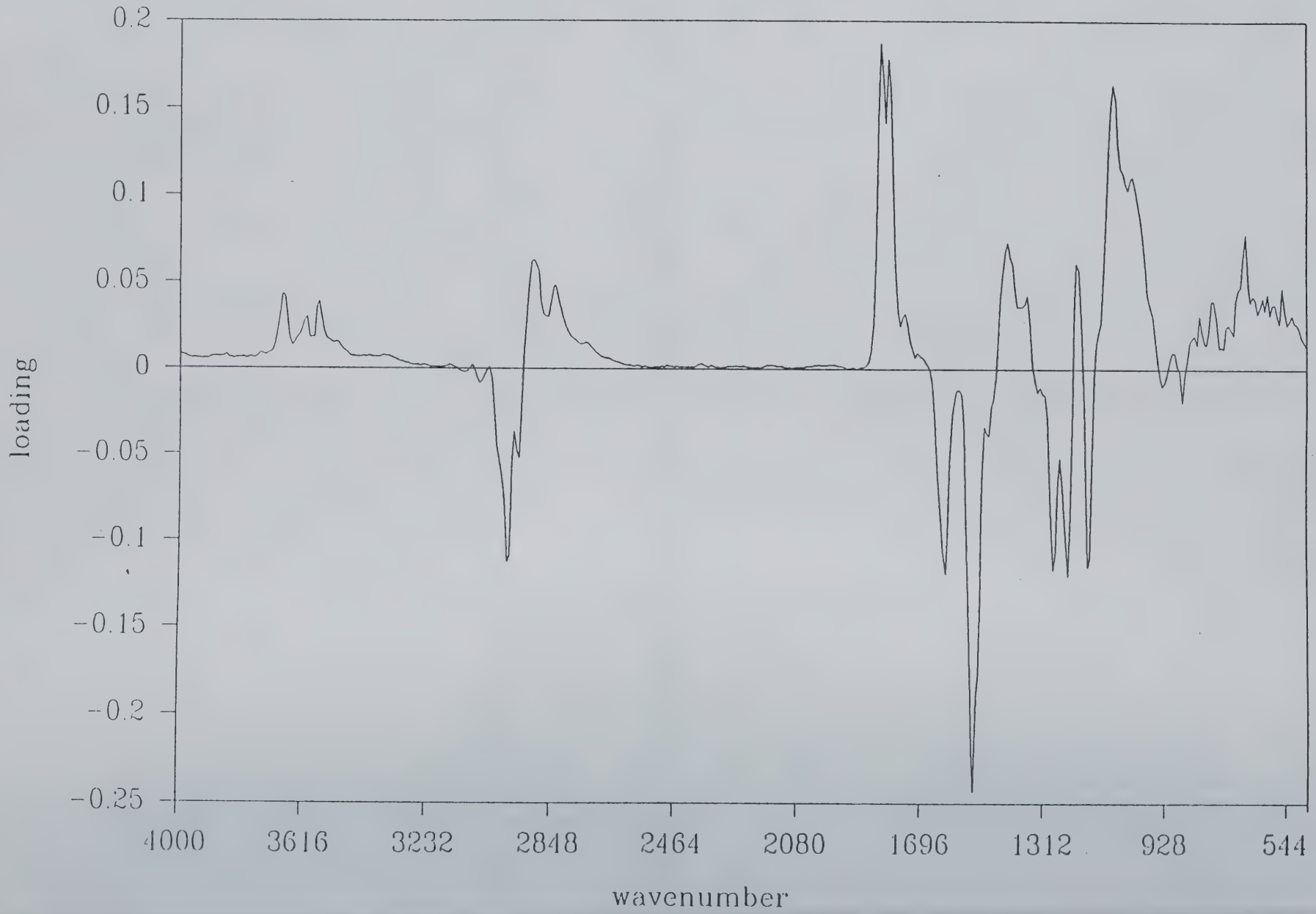


Figure 5 Plot of residual F-norm of the untreated data matrix as principal components are removed.

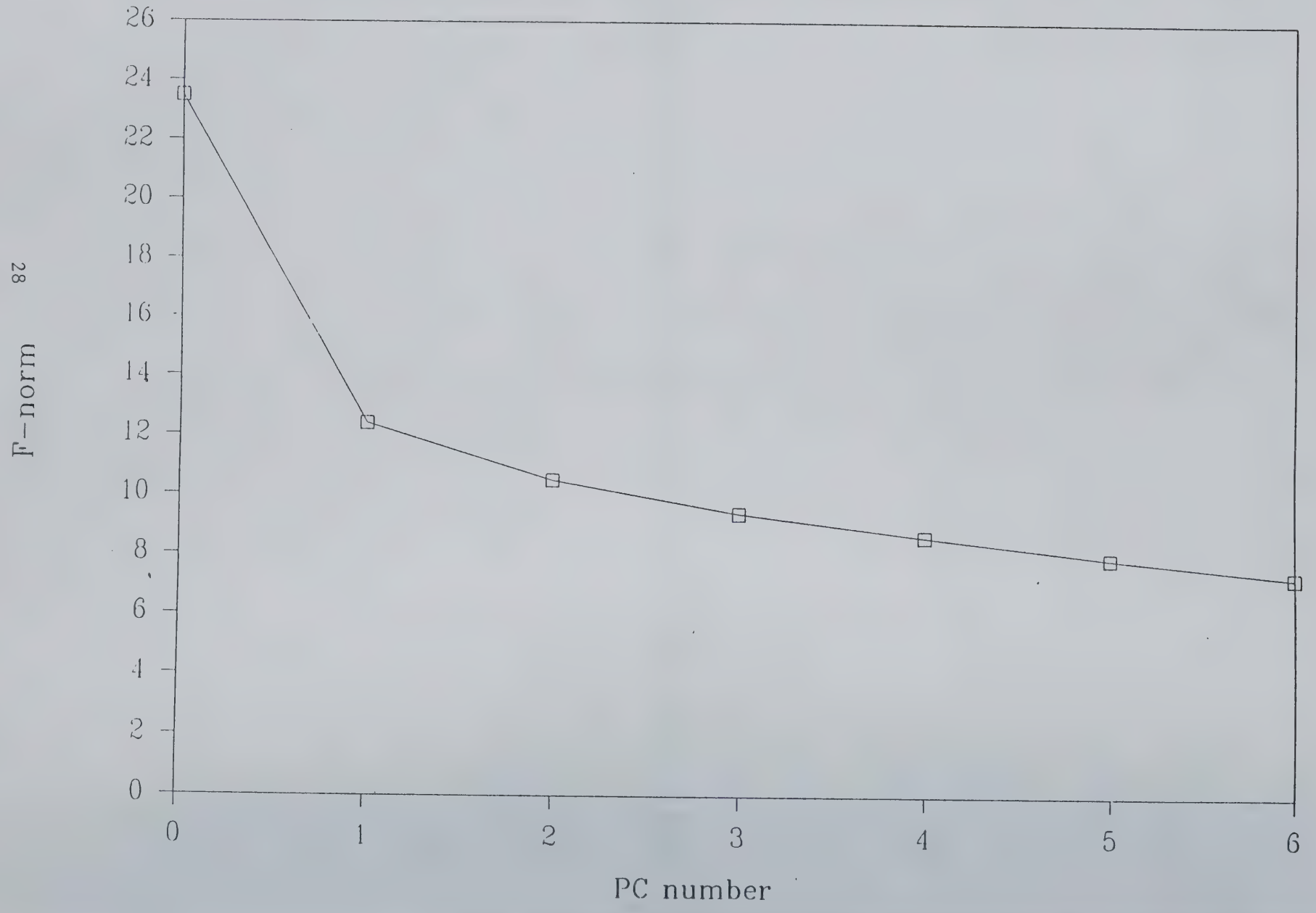

Residual F-norm
F-norm
26
24
22
20
18
16
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
PC number

Figure 6 Mean spectrum of the alcohol/non-alcohol data set.

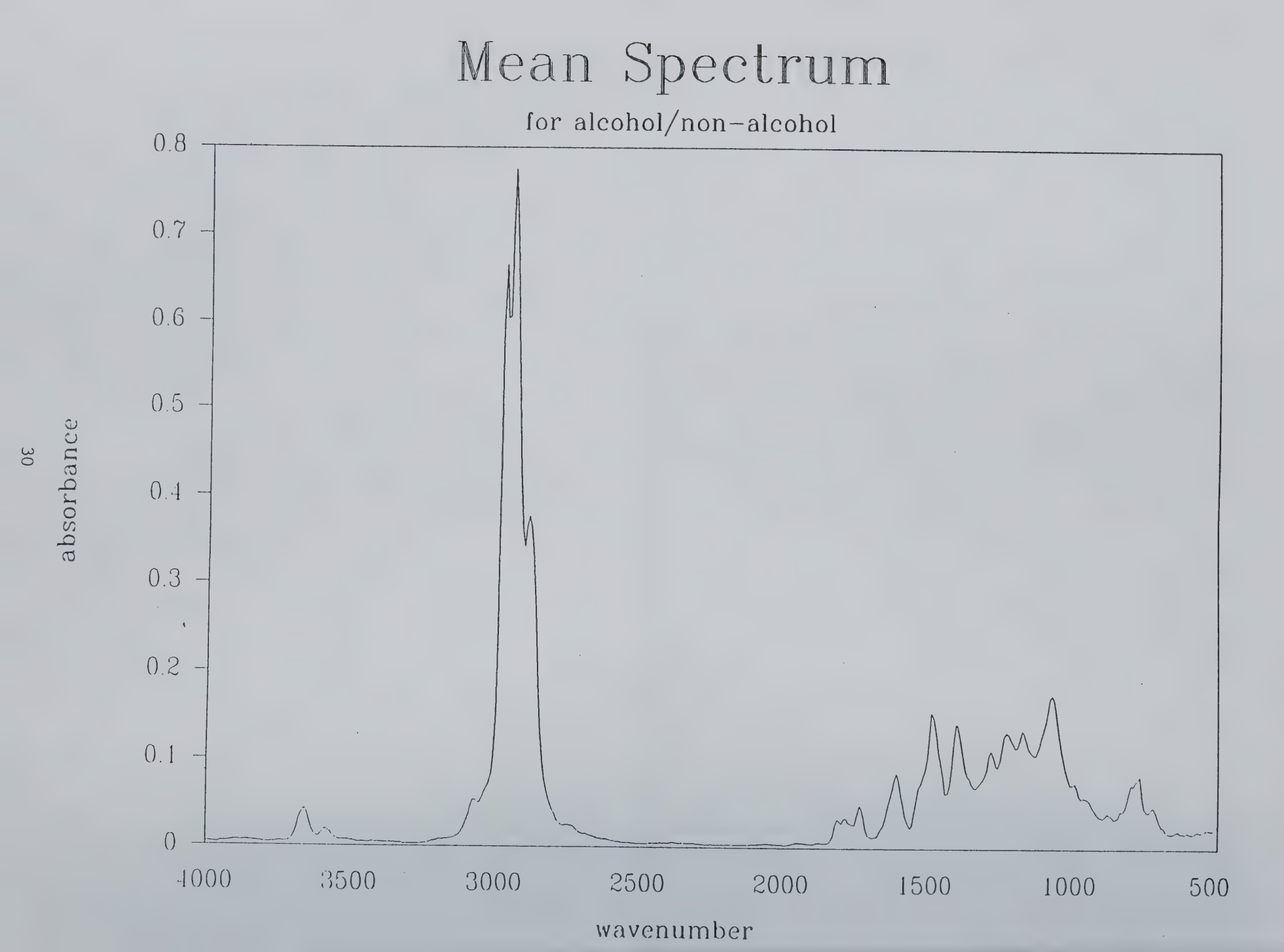

Mean Spectrum
for alcohol/non-alcohol
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
absorbance
4000
3500
3000
2500
2000
1500
1000
500
wavenumber

Figure 7 SQSS spectrum of the alcohol/non-alcohol data set.

# SQSS spectrum

for alcohol/non-alcohol

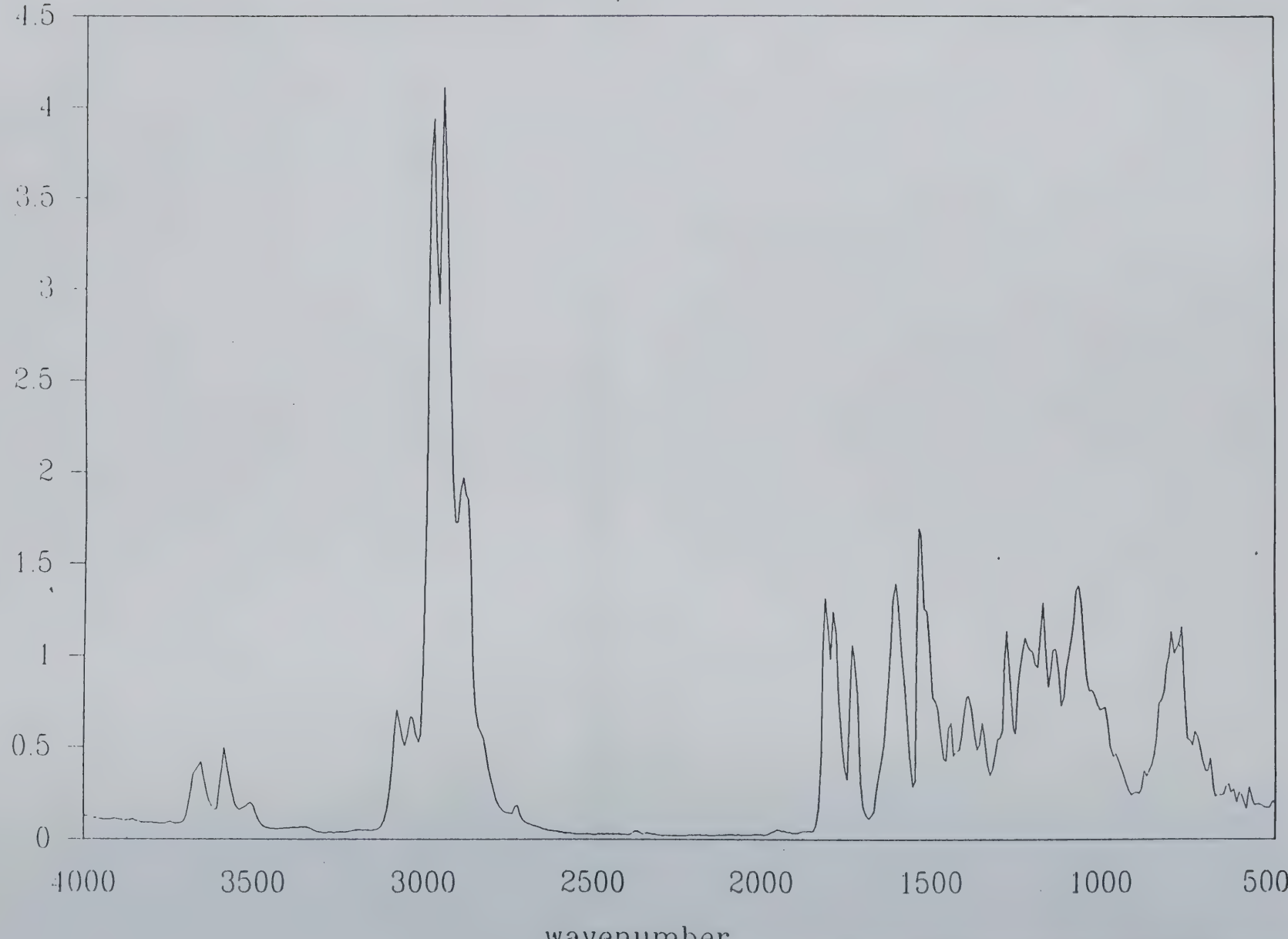


Figure 8 Feature weight spectrum for the alcohol/non-alcohol data set.

# Feature Weights

for alcohol/non-alcohol

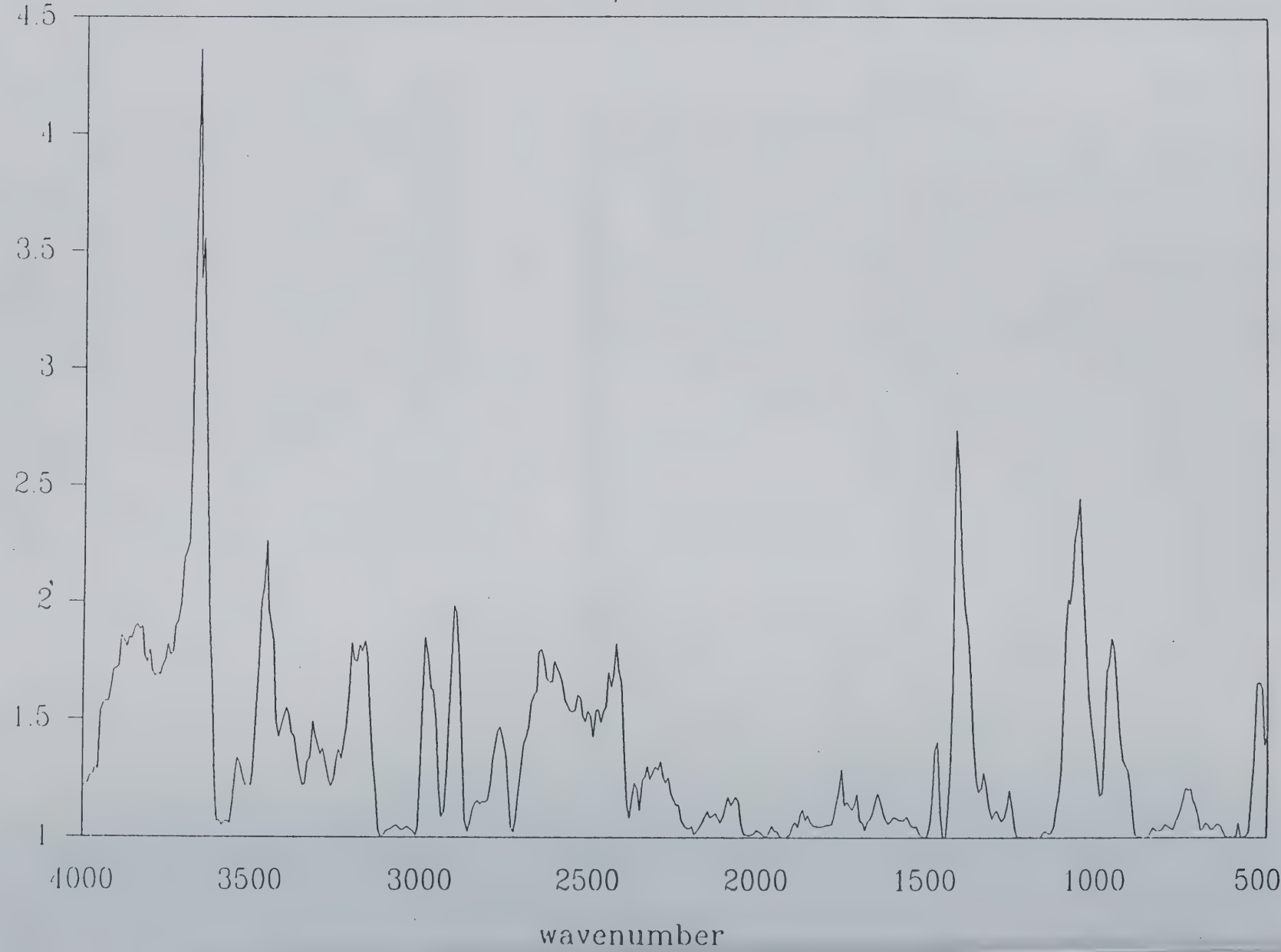


Figure 9 Scores plot for components 1 and 2 for the alcohol/non-alcohol classification with autoscaled and feature weighted data.

# Autoscaled and Feature Weighted Data

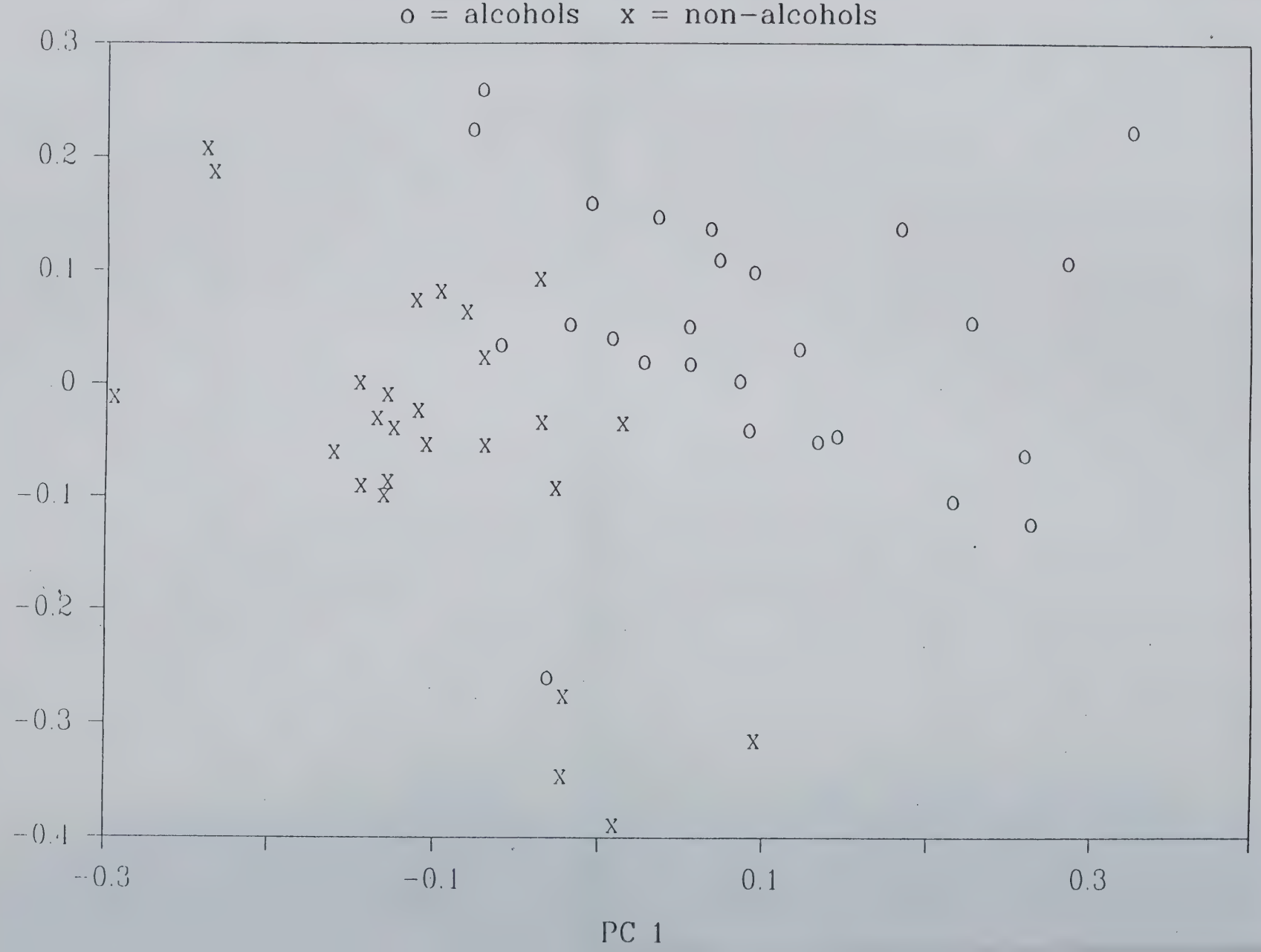


Figure 10 Scores plot for components 1 and 2 for the alcohol/non-alcohol classification with autoscaled and feature weighted data with ellipses set at 1 standard deviation and 2.06 standard deviations (95% confidence limits).

# Autoscaled and Feature Weighted Data

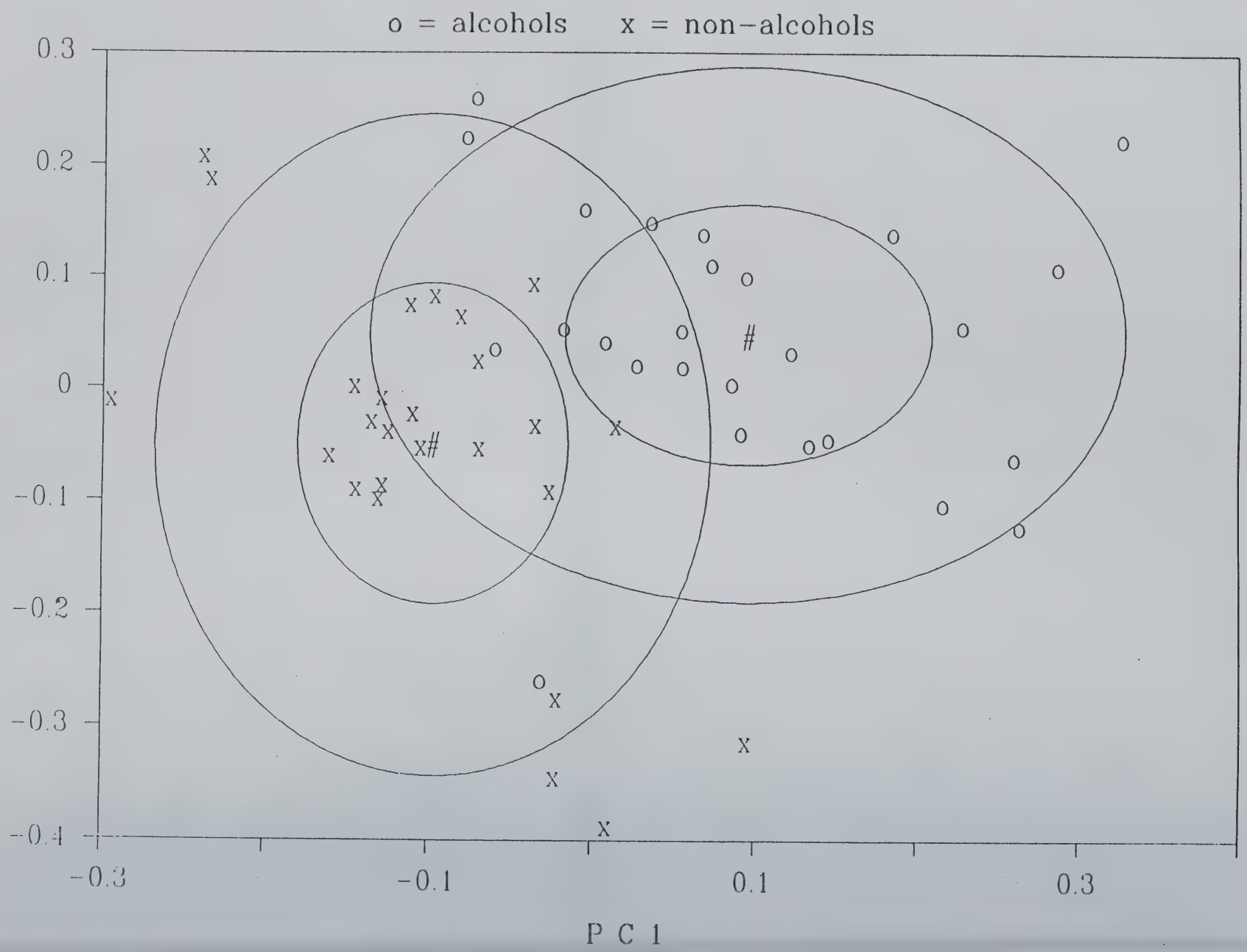


Figure 11 Loading plot for component 1 for the alcohol/non-alcohol classification with autoscaled and feature weighted data.

# PC 1 Loading Spectrum

for alcohol/non-alcohol

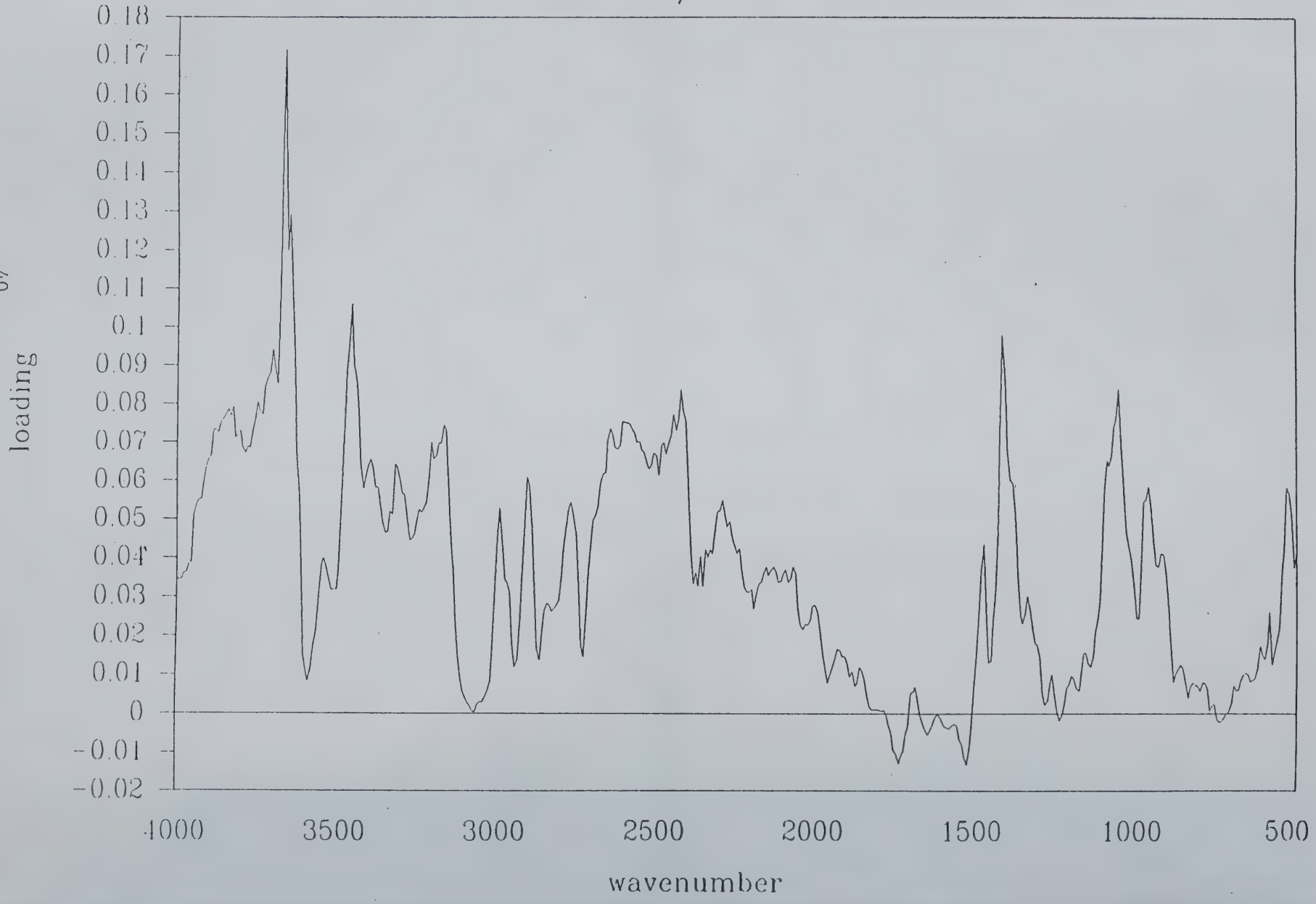


Figure 12 Loading plot for component 2 for the alcohol/non-alcohol classification with autoscaled and feature weighted data.

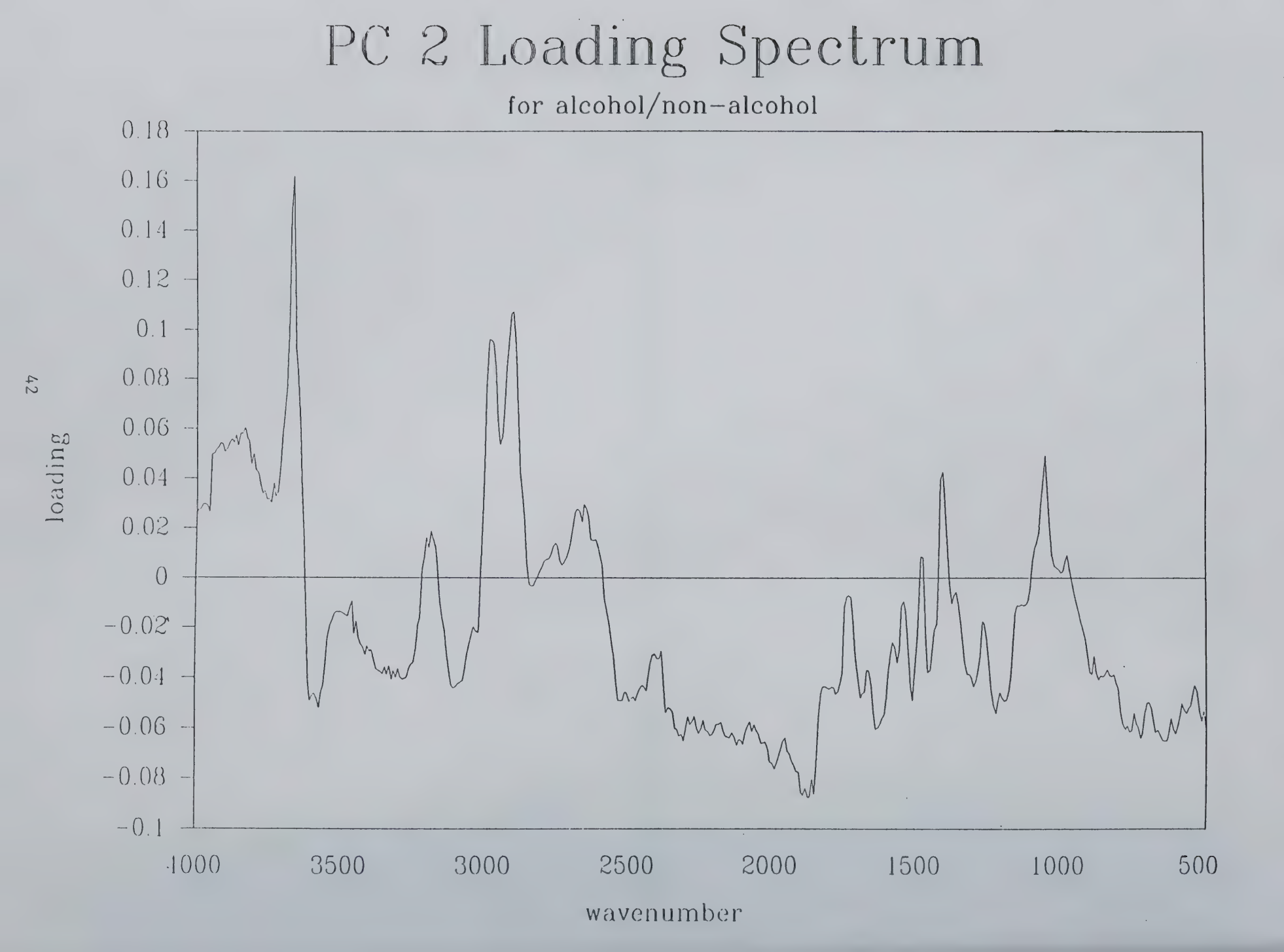

PC 2 Loading Spectrum
for alcohol/non-alcohol
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
loading
4000
3500
3000
2500
2000
1500
1000
500
wavenumber

Figure 13 Squared feature weight spectrum for alcohol/non-alcohol data set.

# Squared Feature Weights for Alcohols

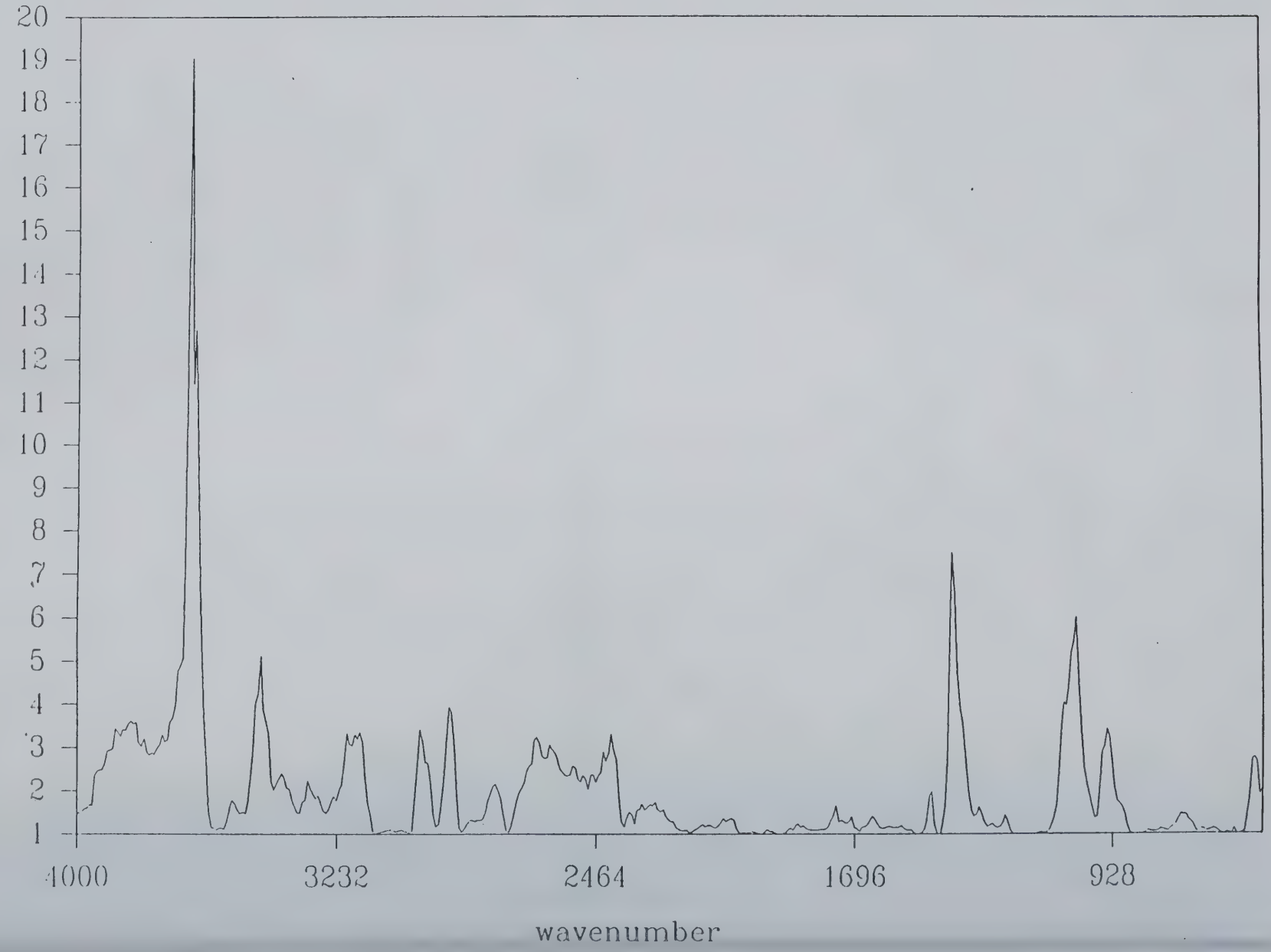


Figure 14 Scores plot for components 1 and 2 for the alcohol/non-alcohol classification with autoscaled and squared feature weighted data.

# Autoscaled Squared Feature Weighted Data

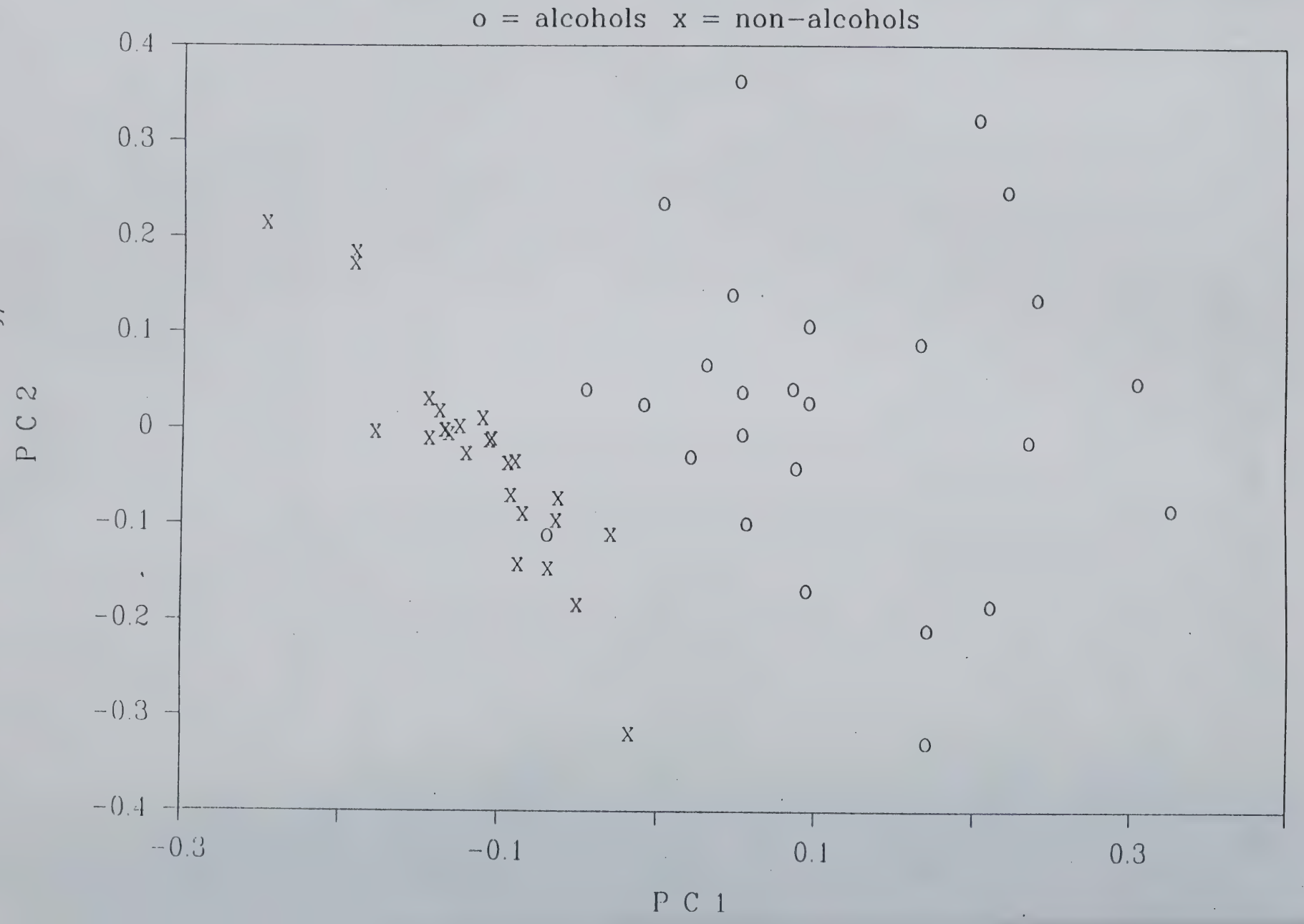


Figure 15 Loading plot for component 1 for the alcohol/non-alcohol classification with autoscaled and squared feature weighted data.

# P C 1 Loading

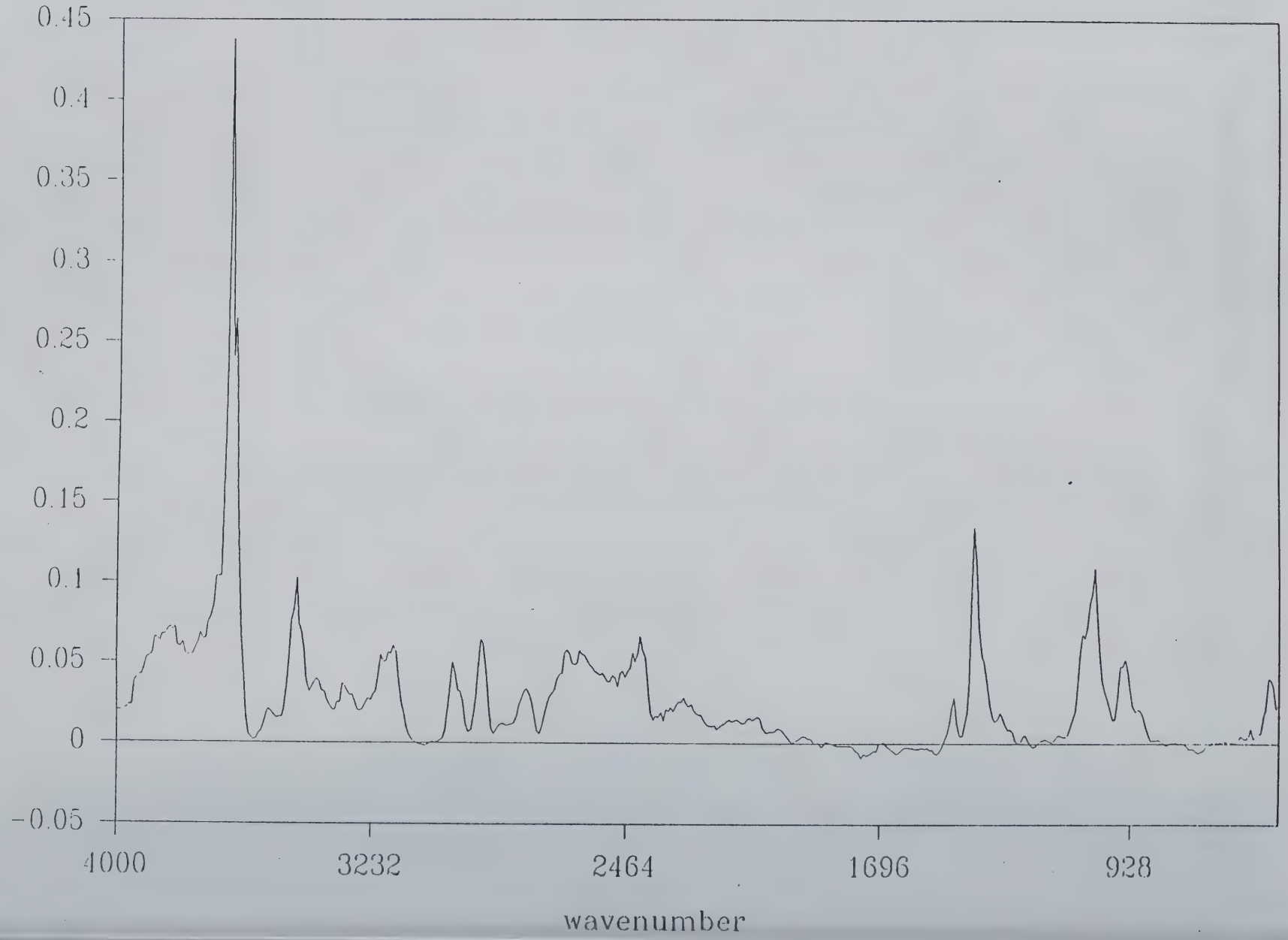

0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
weight
4000
3232
2464
1696
928
wavenumber

Figure 16 Loading plot for component 2 for the alcohol/non-alcohol classification with autoscaled and squared feature weighted data.

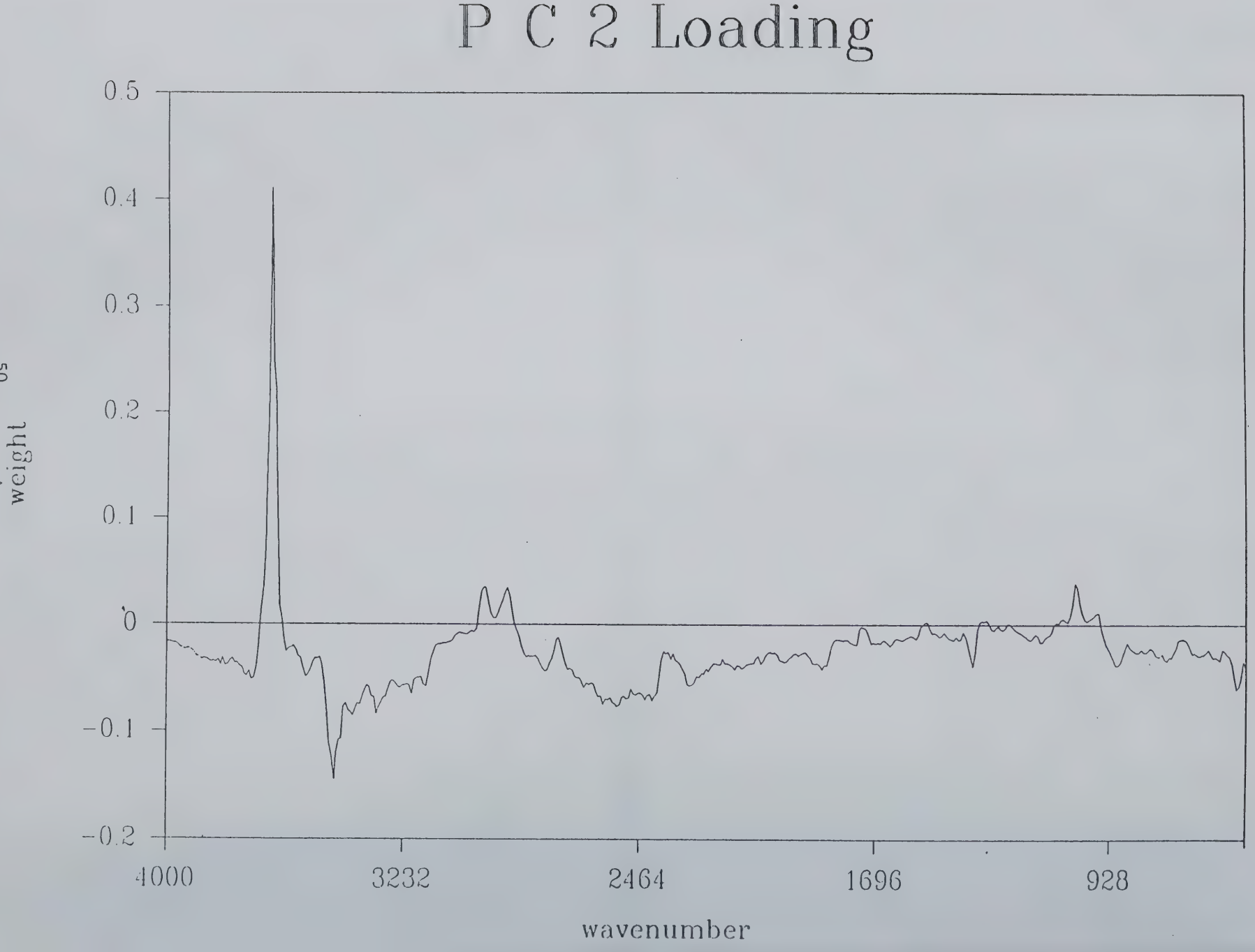

P C 2 Loading
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
weight
4000
3232
2464
1696
928
wavenumber

Figure 17 Mean spectrum of the aromatic/non-aromatic data set.

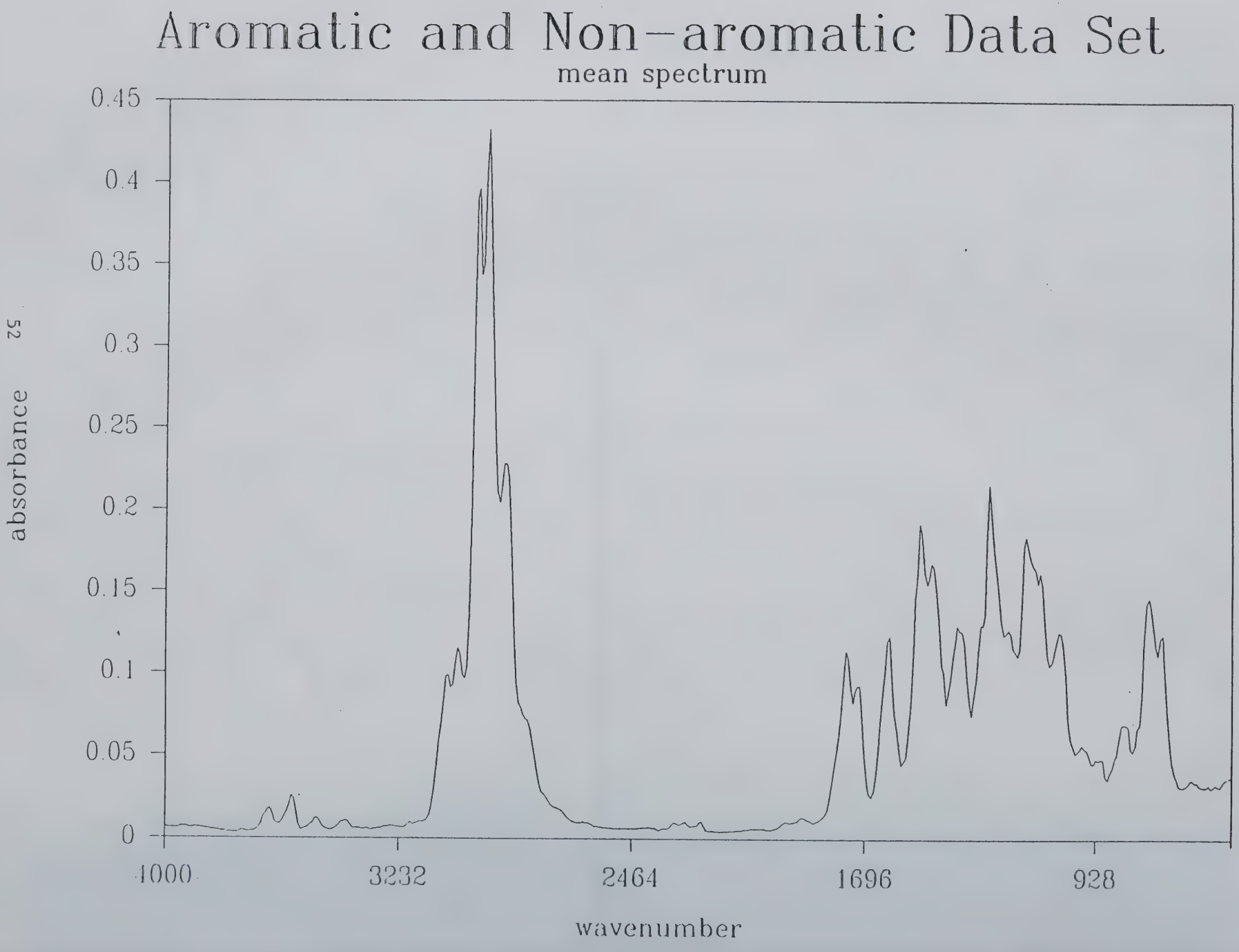

Aromatic and Non-aromatic Data Set
mean spectrum
absorbance
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1000
3232
2464
1696
928
wavenumber

Figure 18 SQSS spectrum of the aromatic/non-aromatic data set.

# Aromatic and Non-aromatic Data Set

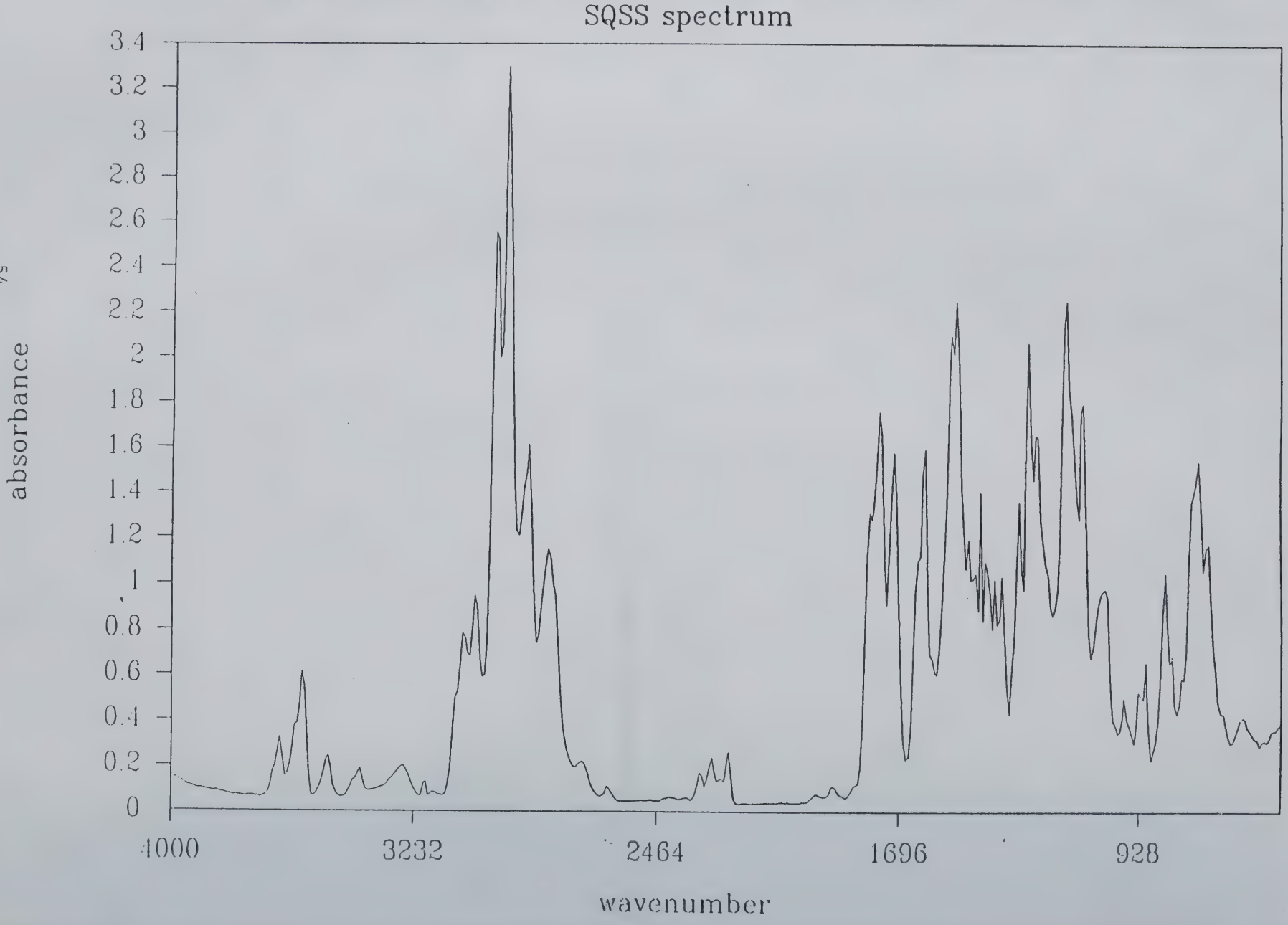


Figure 19 Feature weight spectrum for the aromatic/non-aromatic data set.

# Aromatic and Non-aromatic Data Set

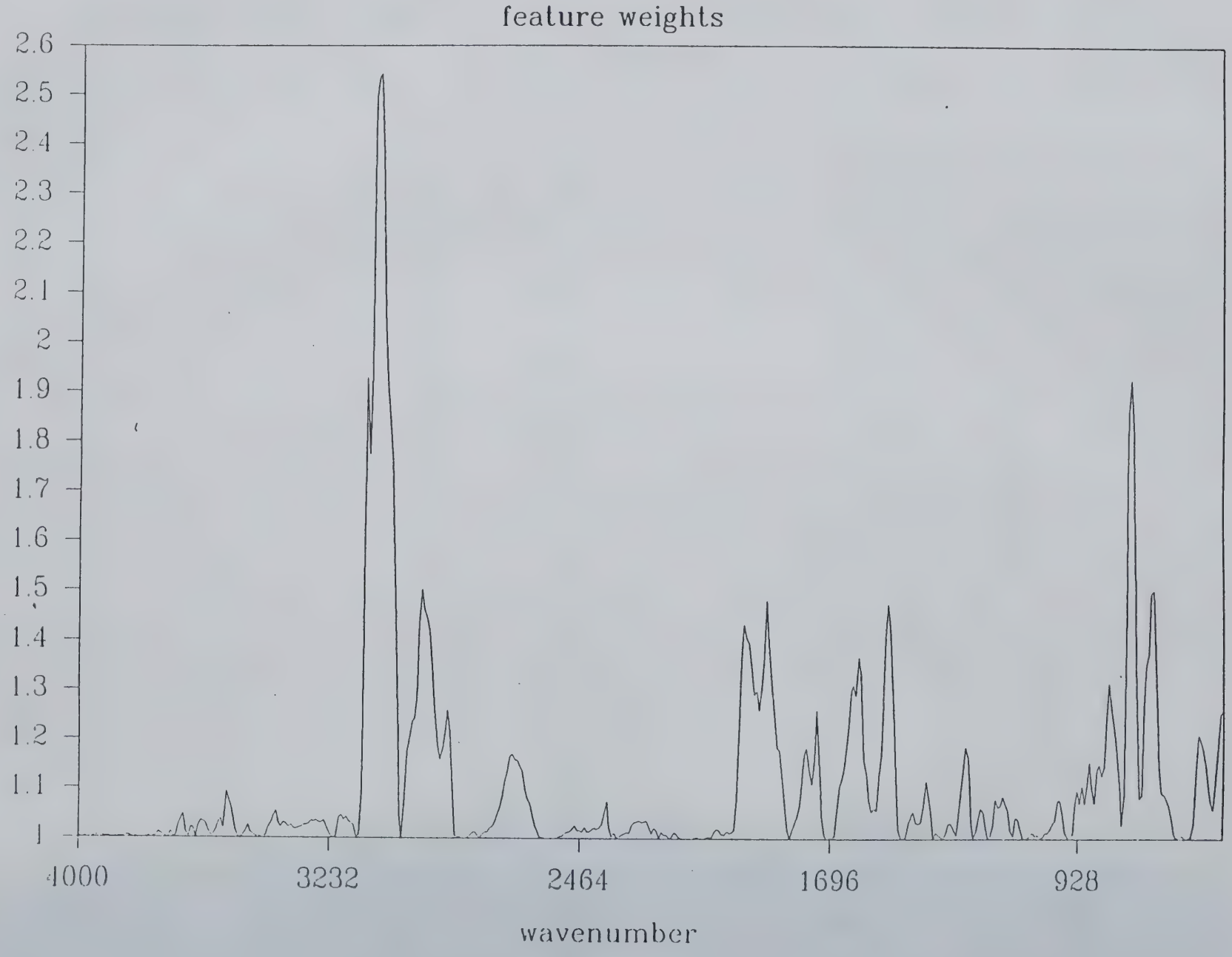


Figure 20 Scores plot for components 1 and 2 for the aromatic/non-aromatic classification with autoscaled and feature weighted data with a separating line drawn between the two classes.

# Autoscaled Feature Weighted Data Set

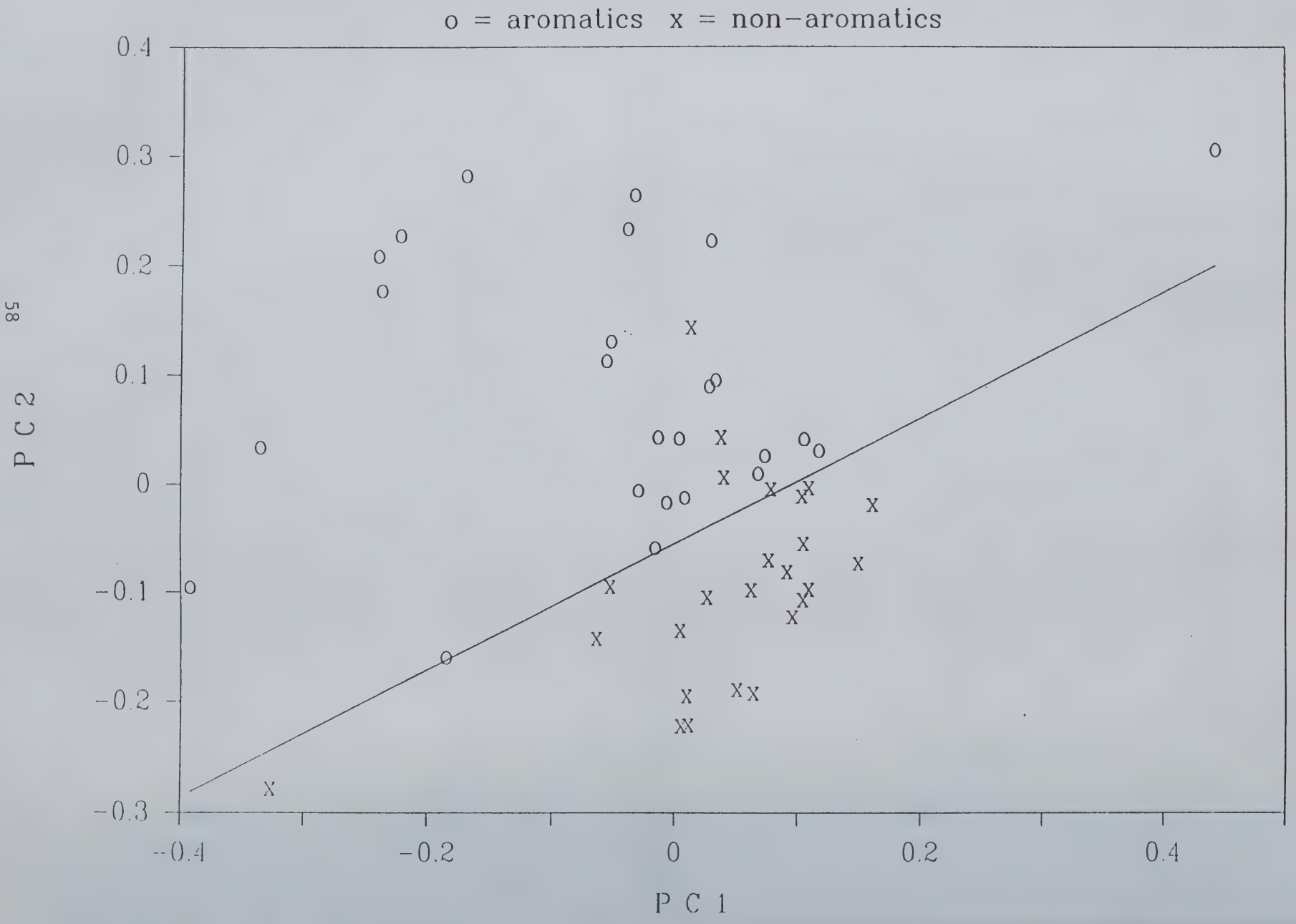


Figure 21 Scores plot for components 1 and 2 for the aromatic/non-aromatic classification with autoscaled and feature weighted data with ellipses set at 1 and 2.06 standard deviations (95% confidence limits).

# Autoscaled and Feature Weighted Data

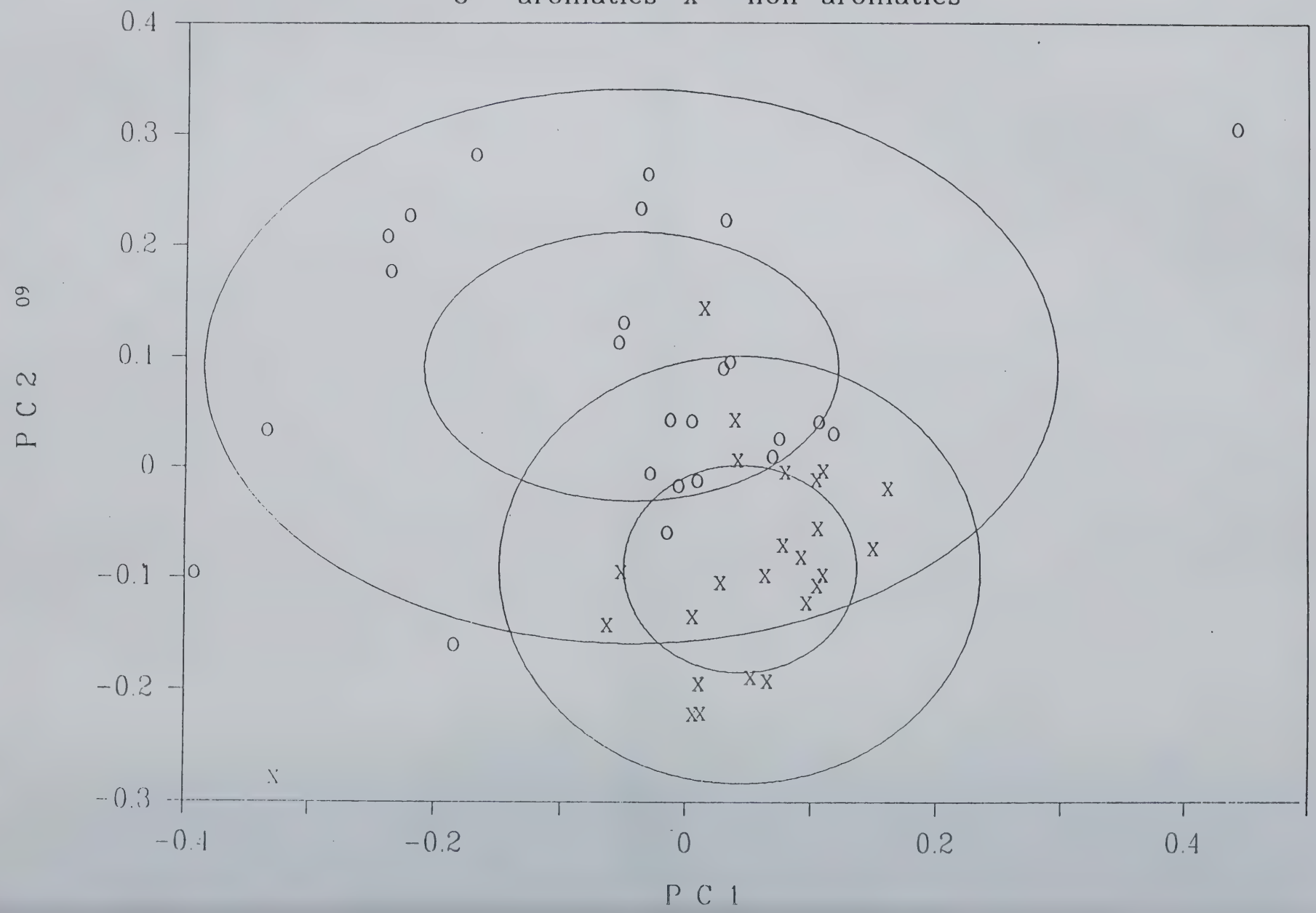


Figure 22 Loading plot for component 1 for the aromatic/non-aromatic classification with autoscaled and feature weighted data.

# Autoscaled Feature Weighted Aromatics

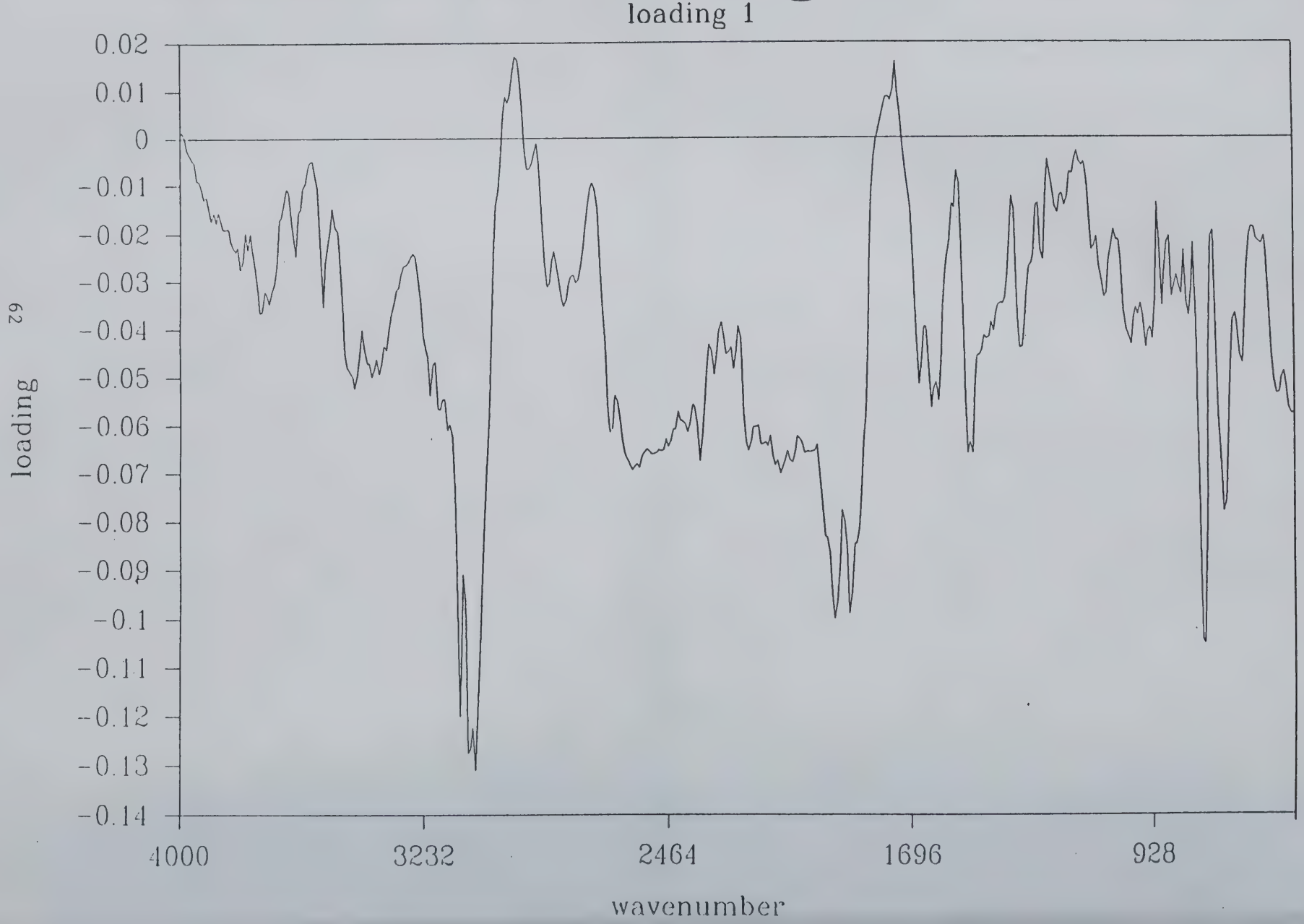


Figure 23 Loading plot for component 2 for the aromatic/non-aromatic classification with autoscaled and feature weighted data.

# Autoscaled Feature Weighted Aromatics

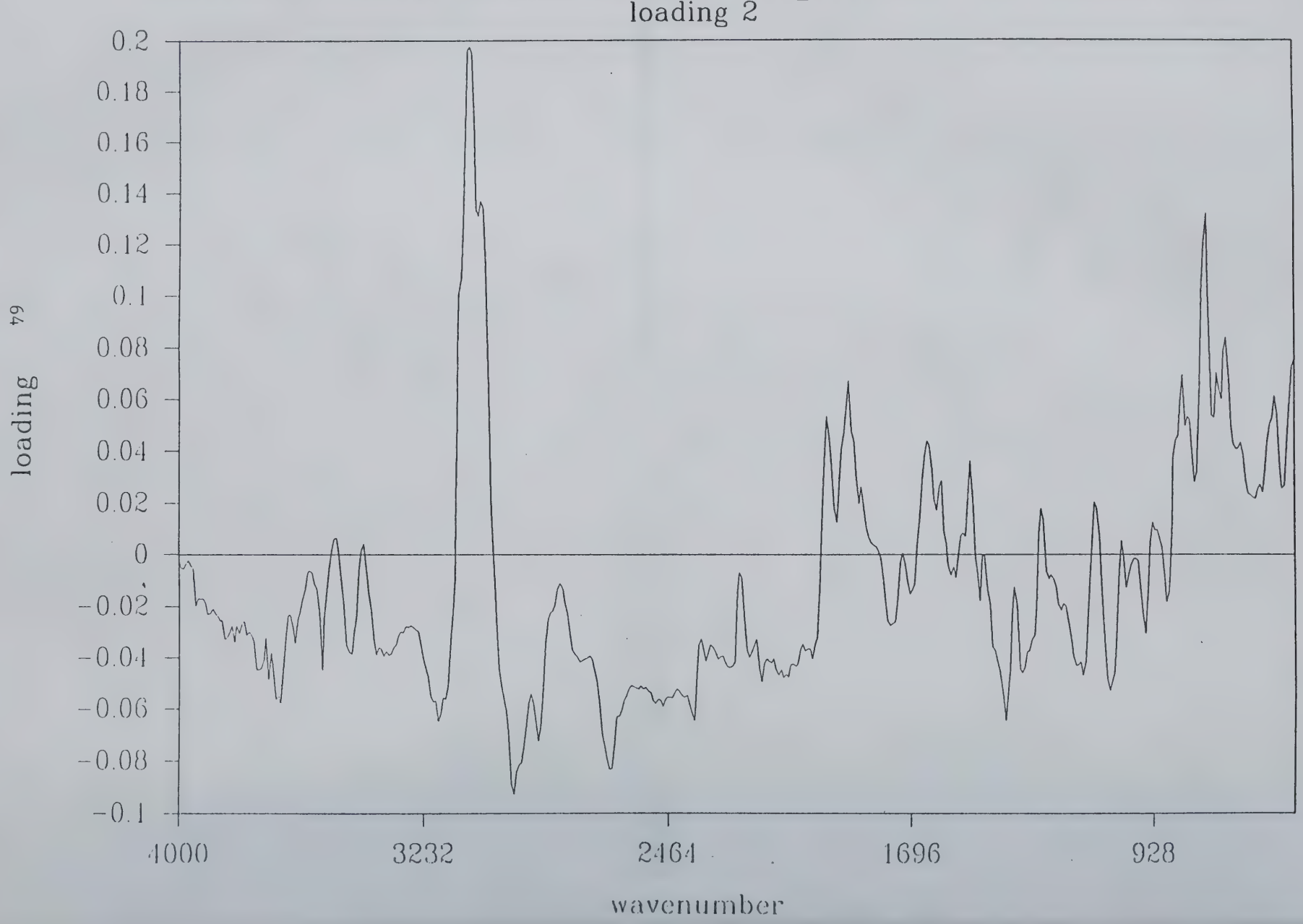


Figure 24 Squared feature weight spectrum for the aromatic/non-aromatic data set.

# Squared Feature Weights

for aromatic/non-aromatic data

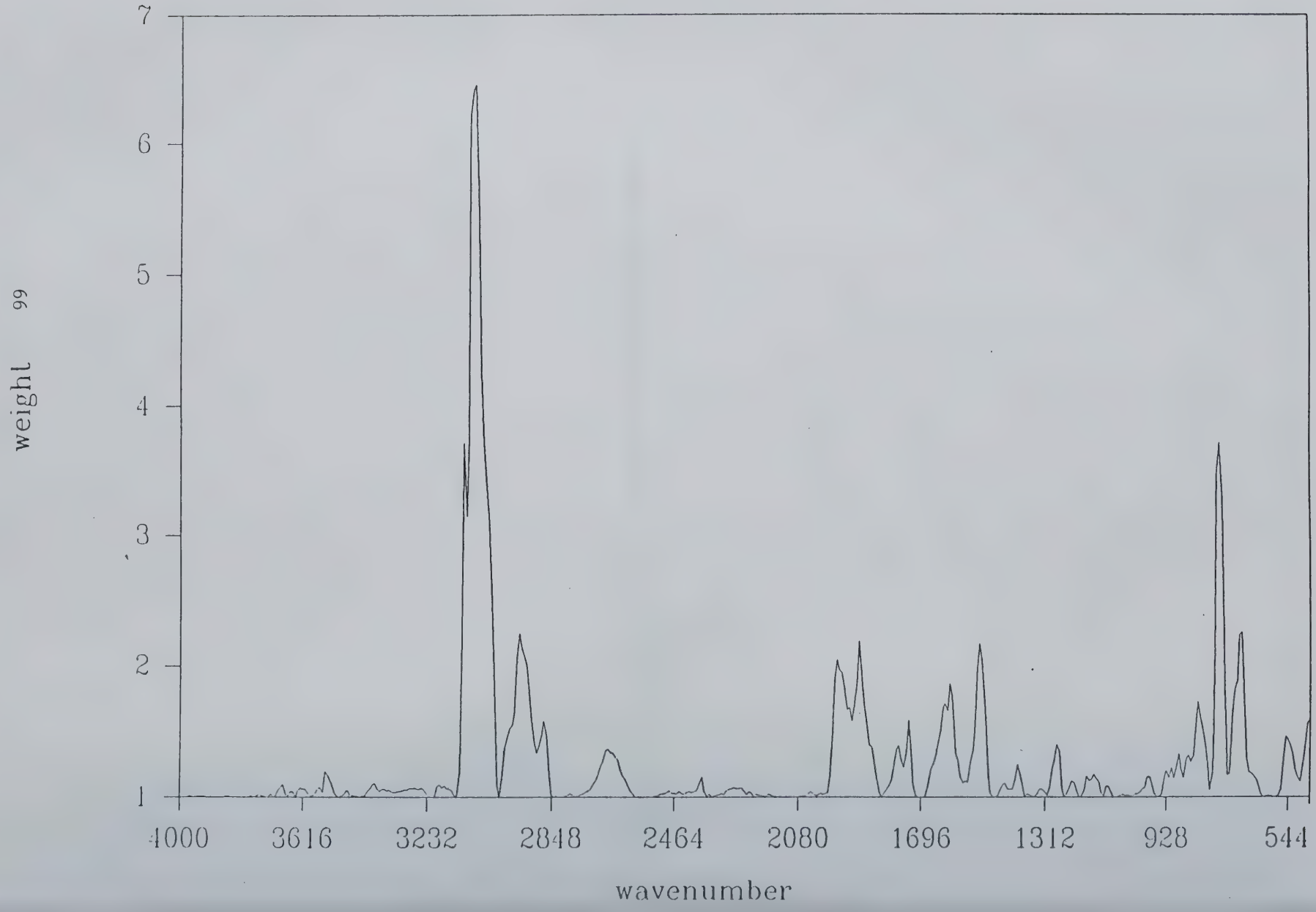


Figure 25 Scores plot for components 1 and 2 for the aromatic/non-aromatic classification with autoscaled and squared feature weighted data.

# Autoscaled and Squared Feature Weighted Data

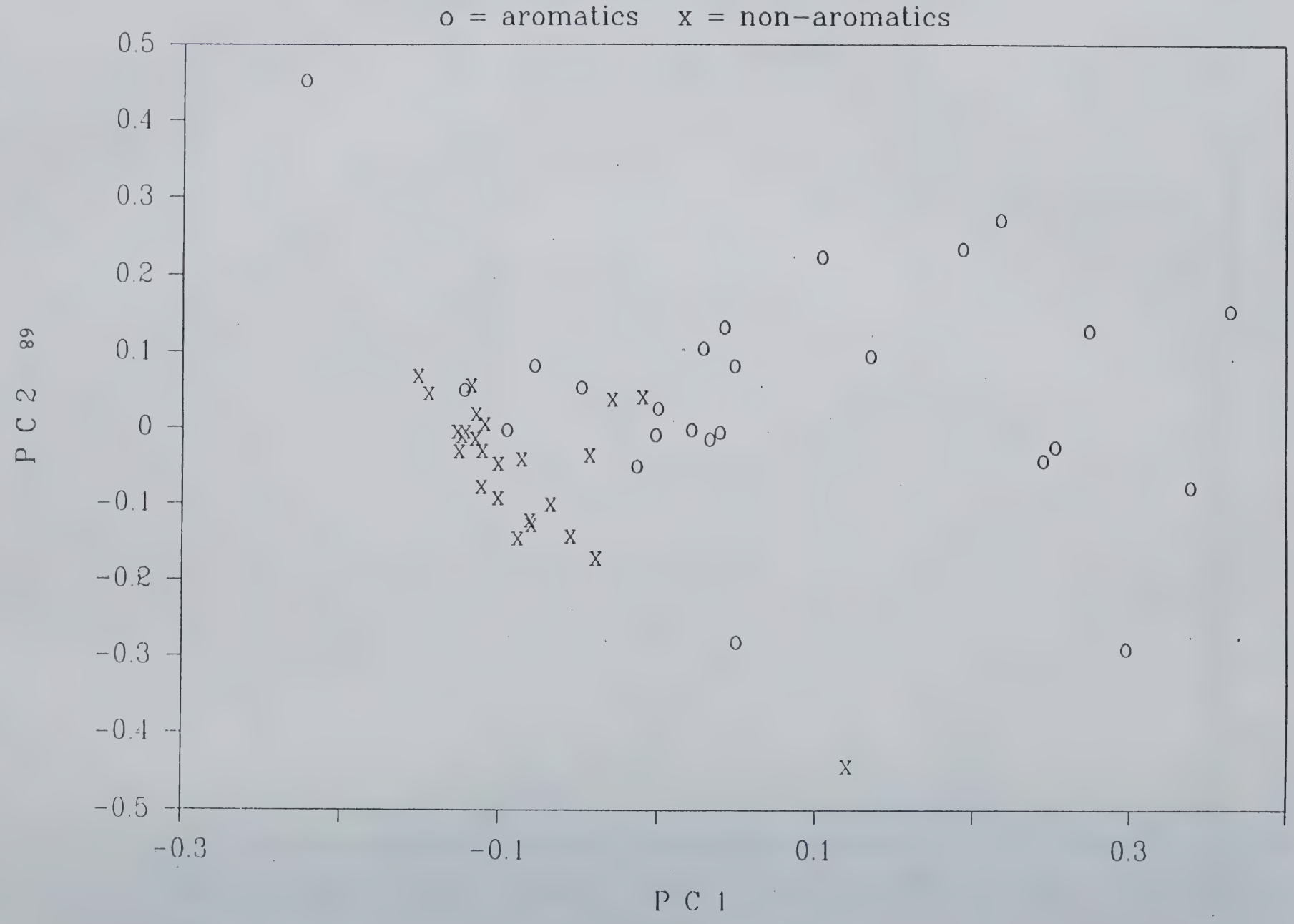


Figure 26 Scores plot for components 1 and 2 for the aromatic/non-aromatic classification with autoscaled and squared feature weighted data with ellipses set at 1 and 2.06 standard deviations (95% confidence limits).

# Autoscaled and Squared Feature Weighted Data

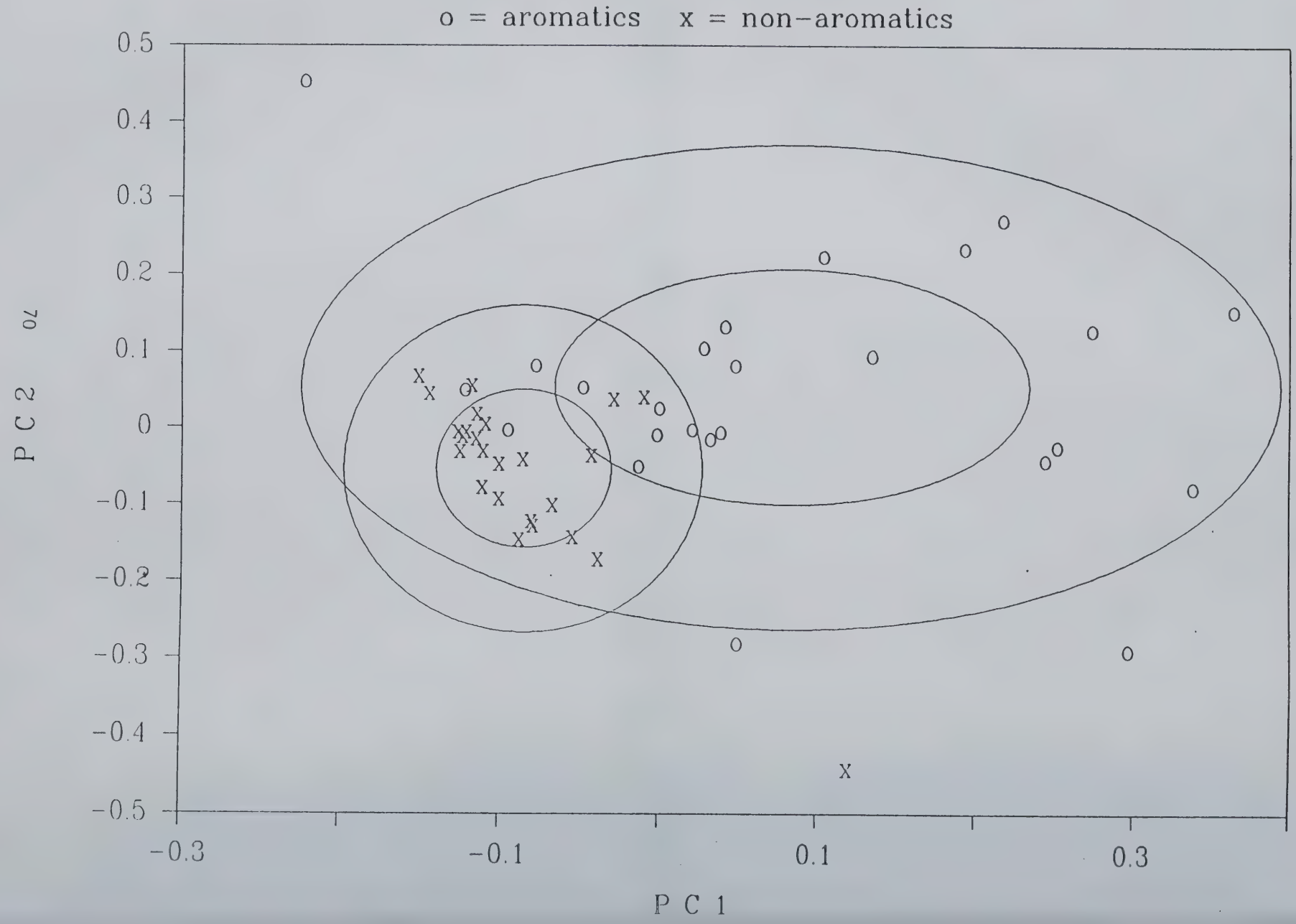


Figure 27 Loading plot for component 1 for the aromatic/non-aromatic classification with autoscaled and squared feature weighted data.

# PC 1 Loading


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
loading
4000
3616
3232
2848
2464
2080
1696
1312
928
544
wavenumber

Figure 28 Loading plot for component 2 for the aromatic/non-aromatic classification with autoscaled and squared feature weighted data.

# PC 2 Loading

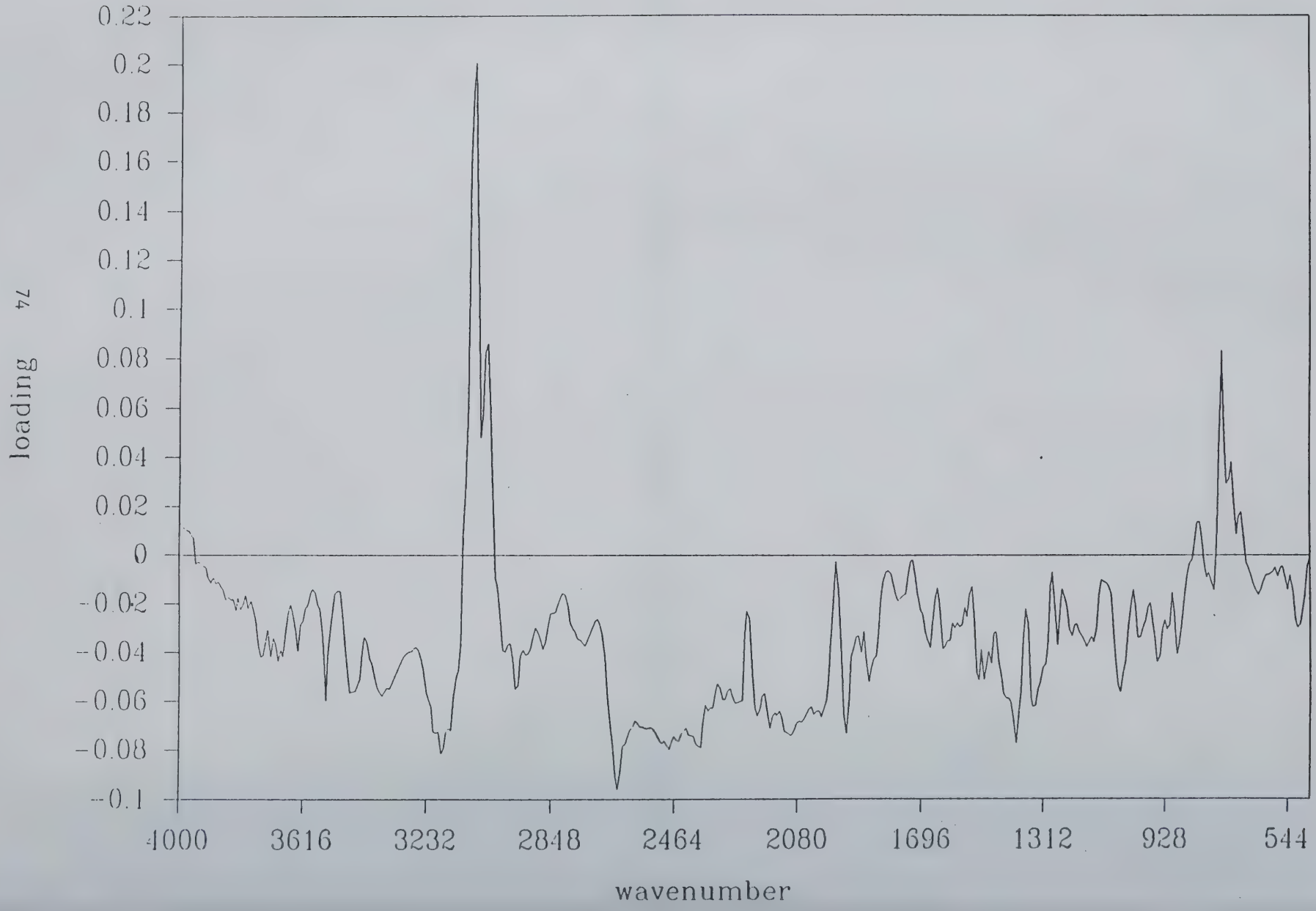

0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
loading
4000
3616
3232
2848
2464
2080
1696
1312
928
544
wavenumber

Figure 29 Scores plot for components 1 and 2 for the alcohol/non-alcohol classification with autoscaled and squared feature weighted data with validation samples.

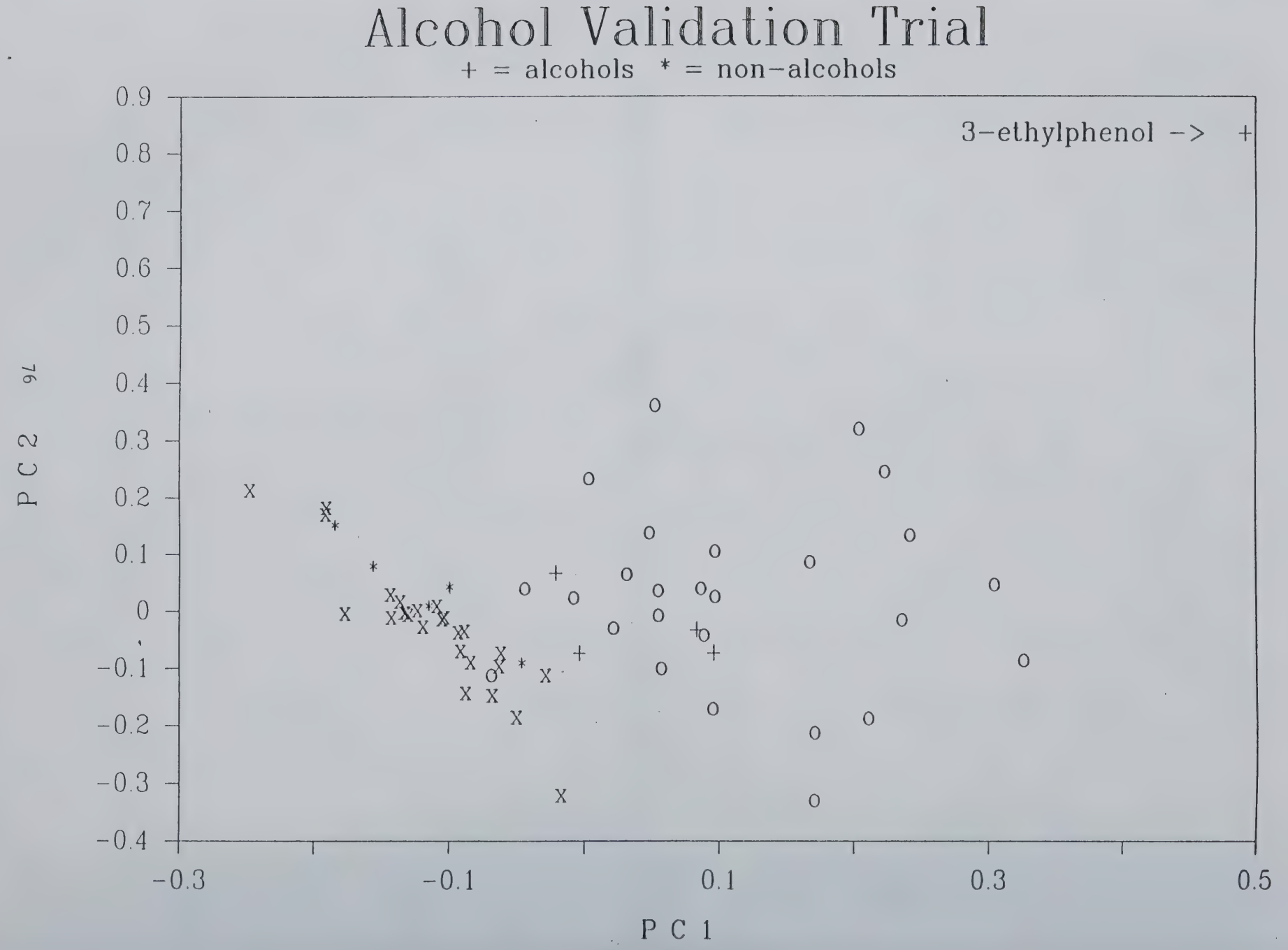
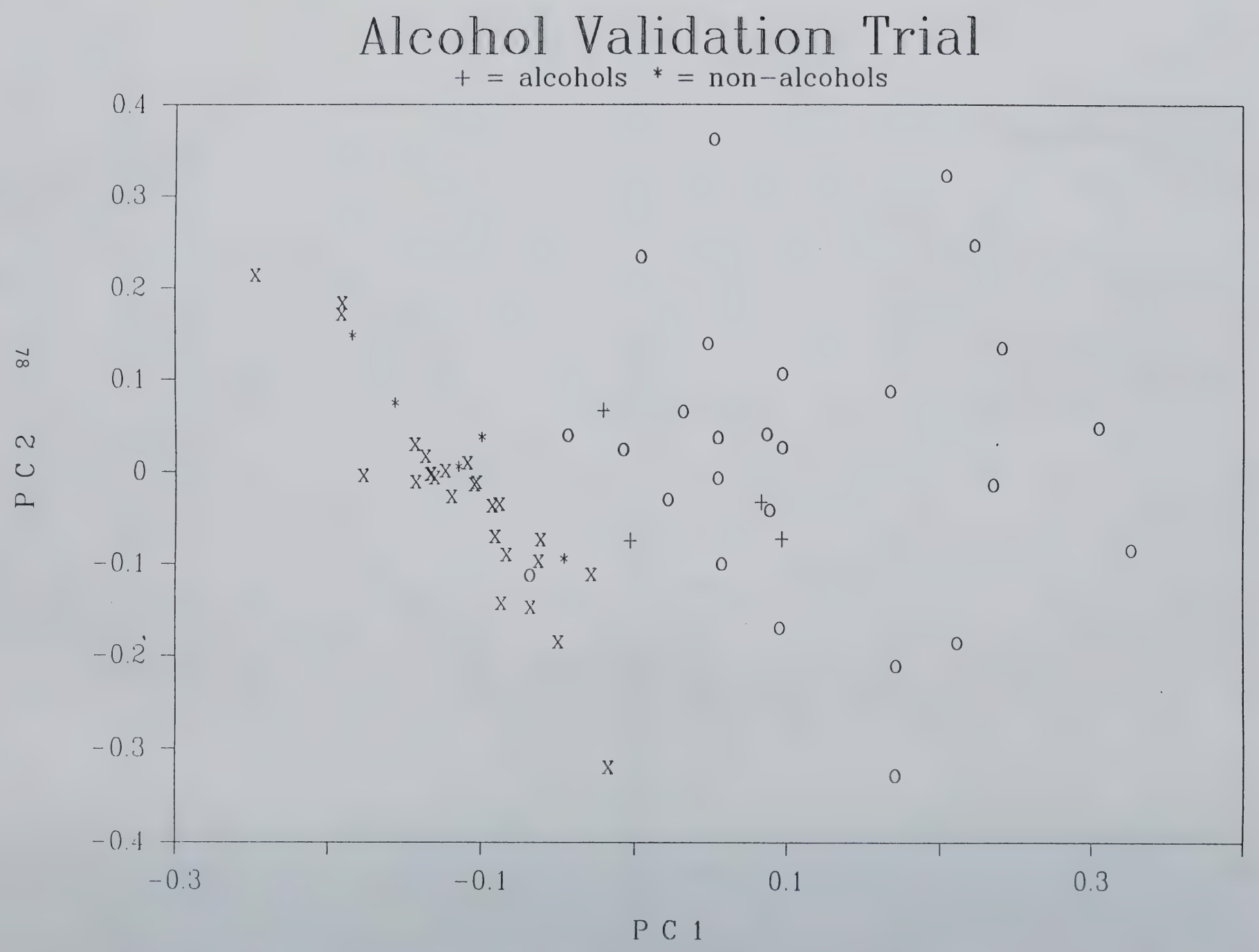

Alcohol Validation Trial
+ = alcohols * = non-alcohols
3-ethylphenol -> +
PC 2 96
PC 1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.1
0.1
0.3
0.5

Figure 30 Expansion of figure 29.


Alcohol Validation Trial
+ = alcohols * = non-alcohols
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.1
0.1
0.3
PC 2
PC 1

NIST-114A
(REV. 3-89)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

# BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NIST-GCR-91-587

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
March 1991

4. TITLE AND SUBTITLE

Expert System Approach for Spectra-Spectra-Structure Correlation for Vapor Phase Infrared Spectra

5. AUTHOR(S)

Peter R. Griffiths

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

University of Idaho
Department of Chemistry
Moscow, ID

7. CONTRACT/GRANT NUMBER
Grant No. 60NANB7D0736

8. TYPE OF REPORT AND PERIOD COVERED
Final Progress 9/1/88-6/30/89

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Gaithersburg, MD 20899

10. SUPPLEMENTARY NOTES

[ ] DOCUMENT DESCRIBES A COMPUTER PROGRAM; SF-185, FIPS SOFTWARE SUMMARY, IS ATTACHED.

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

Identification of unknown gas species can be done-at least approximately-by infrared spectrometric measurements alone, without separation (that is, without gas chromatography) if a sufficiently robust set of algorithms can be developed for automatic search and identification procedures. The PAIRS computer program is an example of such an automated approach. Existing programs, such as PAIRS, however, do not have the ability of identifying a sufficiently large number of gas species to find regular use in areas such as toxicity of fire gases. For such purposes, it was necessary to develop a more comprehensive approach, one which could include every commonly expected functionality. The present report describes the first stage of such development. The solution taken is based on principal components analysis. The report documents the initial development done on this topic.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

Infrared spectrometers, FT-IR, expert systems

13. AVAILABILITY

[X] UNLIMITED

[ ] FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).

[ ] ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.

[X] ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
82

15. PRICE
A05

ELECTRONIC FORM